

硕士学位论文

**面向商品评论情感分析的词相似度的研究与应用**

学位申请人：徐彬

指导教师：曾明 教授

类别（领域）：工程硕士（软件工程）

2017年4月

**Research and Application of Word Similarity for**

**Sentiment Analysis of Commodity Comments**

A thesis submitted to

Xi’an Jiaotong University

in partial fulfillment of the requirements

for the degree of

Master of Engineering

By

Bin Xu

Supervisor: Prof. Ming Zeng

(Software Engineering)

April 2017

**论文题目：面向商品评论情感分析的词相似度的研究与应用**

**类别（领域）：软件工程**

**学位申请人：徐彬**

**指导教师：曾明 教授**

摘 要

网络技术的发展使得人们的日常生活购物行为越来越网络化，在这个过程中当中人们通过对商品的评论来表达对商品的看法。对评论短文本的特性，本文提出了适合进行情感分析的文本表示模型，并针对模型建立过程中的关键部分改进了已有的方法。

本文深入研究了短文本的特点，分析了经典的文本表示模型优缺点，引入词汇链模型和布尔模型的特性，将向量空间模型变化为文本对象空间模型，针对模型建立当中关键的对象抽取和归纳部分，在现有的基于HowNet进行词相似度计算的方法的基础上，改进了词相似度计算。针对基于HowNet的词相似度计算方法无法计算未收录词的缺点，对神经概率语言模型中的CBOW模型进行了改进。同时本文针对词的向量化表示利用关系向量进行相似度计算。本文选择基于HowNet的改进的词相似度计算方法对情感词、程度词以及基于改进的CBOW模型的词向量的词相似度计算方法对名词分别进行对象抽取和归纳。根据短文本的句子成分依存关系分析所得到的对象链计算每个对象维度的值。建立了主题相关的维度选择，在整个文本对象空间当中进行主题相关的维度选择。

实验结果表明，基于HowNet的改进后的方法对于情感词、程度词相似度的计算的准确率提升很大；基于改进的CBOW模型的词向量能够更好的计算名词的相似度，并且采用分类法度量词相似度可以无需人工指定相似度阈值并且具有良好的效果；本文提出的模型所表示的电子商务用户评论短文本能够很好的表达短文本所描述的事物和情感倾向。最后，面向天猫商城用户评论，利用本文模型进行情感分析和商品关注点分析。

**关 键 词**：商品评论；词相似度；对象链；维度选择；情感分析

**论文类型**：应用研究

**Title：Research and Application of Word Similarity for Sentiment Analysis of Commodity Comments**

**Professional Fields：Software Engineering**

**Applicant:Bin Xu**

**Supervisor:Prof.Ming Zeng**

ABSTRACT

With the development of network technology, people are more and more inclined to online shopping in daily life.In this process, people express their opinions by commenting on the goods.According to the characteristics of short text, thesis puts forward a suitable text representation model for sentiment analysis, and puts forward new methods for the key part of the model building process.

According to the the advantages and disadvantages of the classical text representation model,then introduced the features of Lexical chain model and Boolean model to establish text objects space model.In the object extraction and induction part of model building, thesis puts forward improved word similarity calculation on the basis of the existing methods of word similarity calculation based on HowNet and then compare with the existing methods by experiment.In order to overcome the shortcomings that can not process the not include words, thesis improve CBOW model.At the same time, thesis studies the methods of calculating the similarity of words that with distributed representation.The method based on HowNet and the method based on word vector are used to extract and induct object in affective vocabulary and nouns respectively. Build the objects chain by the analysis of the sentence component dependency,then use the objects chain to calculate the value of each object dimension.Establish the topic related dimension select method.

The experimental results show that the calculation accuracy of word similarity in the improved method based on HowNet has been improved,and the word vector that trained by the improved CBOW model can better calculate the similarity of words.The classification measurement is more accurate and more comprehensive than the distance measurement in express the word similarity, and it is superior to the dissimilarity measurement in the calculate difficulty. The text objects space model can better express the focus and the sentiment of the commodity comments.Finally,design a small application based on the text object space model which includes the sentiment analysis and the focus analysis for the Tmall comments.

**KEY WORDS**:Commodity comments;Word similarity;Objects chain;Dimension selection;Sentiment analysis

**TYPE OF THESIS**:Application Research

目 录

[1 绪论 1](#_Toc22117)

[1.1 研究背景及意义 1](#_Toc2635)

[1.2 国内外研究现状 1](#_Toc15518)

[1.3 论文主要研究内容 3](#_Toc4673)

[1.4 本文的组织结构 3](#_Toc1902)

[1.5 本章小结 4](#_Toc20953)

[2 文本分析方法与技术基础 5](#_Toc4150)

[2.1 文本表示模型 5](#_Toc27121)

[2.2 特征选择方法 6](#_Toc5437)

[2.3 常用分类方法 7](#_Toc17682)

[2.3.1 支持向量机 7](#_Toc20185)

[2.3.2 决策树 8](#_Toc6756)

[2.3.3 KNN 9](#_Toc7955)

[2.3.4 神经网络分类器 10](#_Toc13635)

[2.4 知网（HowNet） 10](#_Toc14187)

[2.5 统计语言模型 11](#_Toc3708)

[2.5.1 n-gram模型 11](#_Toc11674)

[2.5.2 神经网络模型 12](#_Toc27775)

[2.5.3 word2vec中CBOW模型 13](#_Toc32646)

[2.6 本章小结 14](#_Toc10754)

[3 基于HowNet的词相似度计算方法 15](#_Toc18376)

[3.1 面向商品评论情感分析的文本表示 15](#_Toc32497)

[3.2 基于HowNet的词相似度计算 15](#_Toc12151)

[3.2.1 基于义原距离和划分的词相似度计算 16](#_Toc17817)

[3.2.2 改进的词相似度计算 18](#_Toc20231)

[3.3 基于HowNet的词相似度计算结果实验分析 21](#_Toc18032)

[3.3.1 义原相似度计算实验结果分析 21](#_Toc32649)

[3.3.2 概念相似度计算实验结果分析 22](#_Toc9968)

[3.3.3 词相似度计算实验结果分析 22](#_Toc26259)

[3.4 本章小结 24](#_Toc15444)

[4 基于词向量的词相似性度量方法 25](#_Toc6791)

[4.1 改进的CBOW模型 25](#_Toc18748)

[4.2 词向量度量词相似性的方法 27](#_Toc27171)

[4.2.1 距离法度量词相似度 27](#_Toc9608)

[4.2.2 分类法度量词相似度 28](#_Toc27004)

[4.2.3 相异法度量词相似度 29](#_Toc2534)

[4.3 基于词向量的词相似度计算实验结果分析 30](#_Toc14283)

[4.3.1 词向量的训练 30](#_Toc17633)

[4.3.2 基于词向量的词相似度计算结果分析 30](#_Toc19645)

[4.4 本章小结 34](#_Toc3456)

[5 基于词相似度的文本对象空间模型的建立和分析 35](#_Toc20769)

[5.1 数据集准备 35](#_Toc1966)

[5.2 文本表示模型的建立 36](#_Toc26153)

[5.2.1 文本对象空间模型 36](#_Toc28926)

[5.2.2 对象的抽取和归纳 37](#_Toc6047)

[5.2.3 词相似度计算方法选择 38](#_Toc24721)

[5.2.4 对象值计算方法 39](#_Toc22704)

[5.2.5 主题相关的对象维度选择 41](#_Toc6947)

[5.3 模型实验结果分析 42](#_Toc10312)

[5.3.1 情感分析评价指标 43](#_Toc17549)

[5.3.2 实验结果分析 44](#_Toc8466)

[5.4 本章小结 46](#_Toc16953)

[6 文本对象空间模型在天猫商品评论分析中的应用 47](#_Toc6798)

[6.1 天猫商城商品评分系统的特点分析与问题 47](#_Toc32351)

[6.1.1 天猫商城商品评分系统的特点 47](#_Toc27625)

[6.1.2 天猫商城商品评分系统存在的问题 47](#_Toc31237)

[6.2 天猫商品评论分析的功能需求 48](#_Toc9844)

[6.2.1 天猫商品评论分析的目标 48](#_Toc19474)

[6.2.2 需求描述 48](#_Toc3306)

[6.2.3 需求分析 48](#_Toc24058)

[6.3 天猫商品评论分析应用的设计 52](#_Toc30192)

[6.3.1 天猫商品评论分析应用的框架结构 52](#_Toc31438)

[6.3.2 天猫商品评论分析应用的模块设计 53](#_Toc31592)

[6.4 天猫商品评论分析应用的模块实现 56](#_Toc11181)

[6.4.1 开发环境 56](#_Toc9619)

[6.4.2 数据层实现 56](#_Toc29937)

[6.4.3 处理层实现 57](#_Toc30745)

[6.4.4 应用层实现 61](#_Toc24969)

[6.5 天猫商品评论分析应用的模块测试 62](#_Toc14443)

[6.5.1 测试环境 62](#_Toc30304)

[6.5.2 功能测试 62](#_Toc16413)

[6.6 天猫商城用户评论分析结果 64](#_Toc13764)

[6.6.1 情感分析 64](#_Toc809)

[6.6.2 关注点分析 65](#_Toc23328)

[6.7 本章小结 66](#_Toc28988)

[7 结论与展望 67](#_Toc31121)

[7.1 结论 67](#_Toc21119)

[7.2 展望 67](#_Toc15392)

[致 谢 69](#_Toc880)

[参考文献 70](#_Toc15512)

声明

CONTENTS

[1 Preface 1](#_Toc27586)

[1.1 Research Background and Significance 1](#_Toc32060)

[1.2 Research Status 1](#_Toc26026)

[1.3 Thesis Content 3](#_Toc16512)

[1.4 Thesis Structure 3](#_Toc29583)

[1.5 Summary 4](#_Toc20442)

[2 Knowledge and Technology Basis for Text Analysis 5](#_Toc23039)

[2.1 Text Representation Model 5](#_Toc8243)

[2.2 Feature Selection Method 6](#_Toc6815)

[2.3 Commonly Used Text Classification Method 7](#_Toc4974)

[2.3.1 Support Vector Machines 7](#_Toc17212)

[2.3.2 Decision Tree 8](#_Toc11729)

[2.3.3 KNN 9](#_Toc26113)

[2.3.4 Neural Network Classifier 10](#_Toc25560)

[2.4 HowNet 10](#_Toc3215)

[2.5 Statistical Language Model 11](#_Toc11110)

[2.5.1 n-gram Model 11](#_Toc17629)

[2.5.2 Neural Network Language Model 12](#_Toc5039)

[2.5.3 CBOW Model in Word2vec 13](#_Toc7685)

[2.6 Summary 14](#_Toc342)

[3 Word Similarity Calculation Method based on HowNet 15](#_Toc23643)

[3.1 Text Representation for Sentiment Analysis of Commodity Comments 15](#_Toc22672)

[3.2 Word Similarity Calculation based on HowNet 15](#_Toc10132)

[3.2.1 Word Similarity Calculation 16](#_Toc27982)

[3.2.2 Improved Word Similarity Calculation 18](#_Toc19136)

[3.3 Experiment and Analysis 21](#_Toc4534)

[3.3.1 Analysis of Semantic Atom Similarity Calculation 21](#_Toc7518)

[3.3.2 Analysis of Concept Similarity Calculation 22](#_Toc7546)

[3.3.3 Analysis of Word Similarity Calculation 22](#_Toc9691)

[3.4 Summary 24](#_Toc23755)

[4 Measurement of Word Similarity based on Word Distributed Representation 25](#_Toc18740)

[4.1 Improved CBOW Model 25](#_Toc3065)

[4.2 Measurement of Word Similarity 27](#_Toc14495)

[4.2.1 Distance Measurement 27](#_Toc27902)

[4.2.2 Classification Measurement 28](#_Toc5465)

[4.2.3 Dissimilarity Measurement 29](#_Toc11918)

[4.3 Experiment and Analysis 30](#_Toc17425)

[4.3.1 Word Distributed Representation Training 30](#_Toc22614)

[4.3.2 Analysis of Experimental Results 30](#_Toc29833)

[4.4 Summary 34](#_Toc20120)

[5 Establishment and Analysis of the Text Objects Space Model 35](#_Toc18277)

[5.1 Corpus Preparation 35](#_Toc20294)

[5.2 Establishment of Text Representation Model 36](#_Toc2874)

[5.2.1 Text Objects Space Model 36](#_Toc7492)

[5.2.2 Object Extraction and Induction 37](#_Toc4963)

[5.2.3 Word Similarity Calculation Method Selection 38](#_Toc29313)

[5.2.4 Object Value Calculation 39](#_Toc32290)

[5.2.5 Topic Related Object Dimension Selection 41](#_Toc22529)

[5.3 Experiment and Analysis 42](#_Toc11536)

[5.3.1 Evaluation Index for Sentiment Analysis 43](#_Toc15119)

[5.3.2 Experimental Results Analysis 44](#_Toc2553)

[5.4 Summary 46](#_Toc4404)

[6 Application of Text Objects Space Model for Tmall Comments Analysis 47](#_Toc17291)

[6.1 The Characteristics and Problems of Tmall Comments system 47](#_Toc10401)

[6.1.1 The Characteristics of Tmall Comments system 47](#_Toc18560)

[6.1.2 The Problems of Tmall Comments system 47](#_Toc28973)

[6.2 Function Analysis for Tmall Comments Analysis 48](#_Toc15669)

[6.2.1 Target of Tmall Comments Analysis 48](#_Toc10786)

[6.2.2 Requirements description 48](#_Toc18114)

[6.2.3 Requirement Analysis 48](#_Toc858)

[6.3 Design for Tmall Comments Analysis 52](#_Toc8212)

[6.3.1 Framework 52](#_Toc29838)

[6.3.2 Module Design 53](#_Toc20634)

[6.4 Implementation for Tmall Comments Analysis 56](#_Toc25531)

[6.4.1 Development environment 56](#_Toc22906)

[6.4.2 Data Layer Implementation 56](#_Toc15367)

[6.4.3 Processing Layer Implementation 57](#_Toc5493)

[6.4.4 Application Layer Implementation 61](#_Toc24726)

[6.5 Module Test 62](#_Toc10465)

[6.5.1 Test Enviroment 62](#_Toc806)

[6.5.2 Function Test 62](#_Toc2649)

[6.6 Results of Tmall Comments Analysis 64](#_Toc25708)

[6.6.1 Sentiment Analysis 64](#_Toc12826)

[6.6.2 Commodity Focus Analysis 65](#_Toc29839)

[6.7 Summary 66](#_Toc11559)

[7 Conclusions and Suggestions 67](#_Toc15403)

[7.1 Conclusions 67](#_Toc31399)

[7.2 Suggestions 67](#_Toc31432)

[Acknowledgements 69](#_Toc3357)

[References 70](#_Toc3493)

Declarations

中看不见即可）：

# 绪论

## 研究背景及意义

随着互联网信息技术的发展，网络购物已经成为一项非常普遍且成熟的交易行为。根据中国互联网信息中心发布的第39次《中国互联网发展状况统计报告》[1]显示，截至 2016年12月，中国网民规模达到了7.31 亿，互联网普及率为 53.2%。网络应用增多和网络普及带来的是一个信息爆炸的时代，人们现在已经开始习惯于通过网络来获得自己想要的信息，但是网络信息更新的速度是极为快速的，如何在大量的信息当中找到对自己有用的成为了一个难题。用户希望可以知道以前的买家对商品的真实评价是负面的还是正面的；社交平台上用户希望能够找到关心的内容和志同道合的潜在好友；网民群体的言论归类成为政府了解舆情和制订政策的依据。

这些越来越迫切的需求，都涉及到一项非常关键的技术，就是文本分析，在上述场景下更确切的说是短文本分析。文本的分类是文本分析方法的一种，分类是一种行之有效的缩小信息规模的方法，采用监督学习的方法在人工标注的数据集上进行训练得到分类器；或者直接将数据集采用聚类的方式自动聚集成簇，再分析每个簇的特征。网页分档、各种长文本通常采用比较主流的机器学习方法如决策树、朴素贝叶斯、支持向量机、k-means等，在短文本分类当中使用这些方法效果并不太好。因为短文本具有如下特点：首先文本长度很短，目前一篇微博最高限制140个字，电子商务用户评论每条也不过几句话，属于每一个短文本的信息单元是少量的。其次短文本涉及的词语总量是巨大的，所以用传统的方法得到的文本的表示维度是非常多的，而每一个短文本放在传统模型当中后，有效维度的占比非常的小，在用传统的方法进行分析时，等于是放大了无关维度的影响因子。这些因素导致适用于长文本的方法在短文本分析方面力有未逮。

一个适合对电子商务评论短文本进行情感分析的模型的建立是值得研究的，考虑单个短文本当中词量少但是词典大的情形，在对短文本进行模型化表示时可以利用词相似度来进行降维。

## 国内外研究现状

在长文本分类当中，大多数经典的算法都是将文本用向量空间模型表示，通常将词语作为空间的维度，词频信息作为维度的值，这样的方法应用到短文本分析当中时，会忽略掉词语本身所具有的描述性信息、词与词之间的相关性信息、一义多词的现象等，所以有必要针对短文本的分析进行专门的研究，建立合适的模型。常用方法是特征扩展，有两种方向：一是通过短文本的隐含主题的挖掘，通过主题来扩展特征；二是通过外部知识库如知网、维基百科、wordnet等或者搜索引擎来进行特征扩展。短文本经过特征扩展之后就可以采用经典的长文本的分析方法进行分析。特征扩展主要在于扩展后的特征要在原本的短文本当中隐含或者相关，有不少的研究的利用知网（中文）和wordnet（英文）当中所呈现的词关系进行语义抽象和特征扩展。有人向量空间模型的基础上将关键词映射到知网当中，抽取特征，然后将这些特征作为关键词在空间中的替代，从而进行分析。陈现麟利用知网中义原的上下位关系扩展到词语层级，从而进行特征扩展[2]；孙建旺使用KNN对基于知网的词的相似度，利用最大匹配方法对文本的相似性进行分类[3]。Zhan采用语言网络的方法从百度百科当中寻找相关特征[4]，语言网络是一个无向图，文本当中的词语是图节点，当中两个词语有关系时存在一条边。Wang先对短文本进行概念化的表示然后进行分类，其概念化的表示来源于使用Probase学习所得的每一个类别的概念化模型[5]。Kim的等人基于维基百科扩展词表[6]。Meng等人将短文本试做一个短字符串，将搜索引擎的结果作为背景知识来扩展特征[7]。Dai提出采用相关性的短文本的聚类使得一堆短文本看做同一个主题，利用主题生成模型使得每个主题的多个特征词构成一个长文本，然后用长文本代替原本的短文本作为分析源[8]。Chen提出通过人工设定LDA主题的数量，得到训练集上不同粒度的主题分布，最后将短文本自身的词频信息与多个粒度上的主题分布拼接作为最终的特征[9]。刘帅等利用条件随机场，在大的语料库里进行特征扩展[10]。

文本的情感分析是采用某种方法自动化的给出某一文本表达的情感是正向或者负向，情感分析关注的是文本当中的观点、喜好和感受，简单的情感分析可以看做是一个二分类。Jaap Kamps基于WordNet下词汇之间的路径关系进行词汇情感倾向的判定[11]；Wiebe提出了词汇在不同的语境当中的倾向性是有区别的，这种区别由词汇语义的相似或者不同来表示[12]。朱嫣岚提出了中文语句的倾向性来源于词汇语义倾向，词汇的语义倾向根据知网来进行计算[13]。孙莹在知网的基础上重新构建了情感词词典，在短文本当中取得了更好的结果[14]。

词相似度的研究有根据词汇语义的方法和根据标注语料库的方法。典型的基于词汇语义的方法是利用预定义的语义词典来进行词相似度的比较，典型的语义词典有针对英文的WordNet、针对中文的HowNet、同义词词林等。基于标注语料库是将词语的语境信息、场景信息等作为词相似度计算的可利用信息。国外Agiree利用WordNet中的词之间的节点路径关系，同时考虑了词语树的深度和网络密度来计算英语词的相似度[15]。P.Brown等人利用大规模语料当中的互信息来进行相似度计算[16]。国内李素建等人提出综合利用HowNet和“同义词词林”来计算词的相似度[17]；李茹等利用框架语义来判断句子相似度，将词相似度的结果隐含在句子相似度结果中[18]。詹志建利用语言网络来对词语的相关集合进行扩展，利用相关集合的近似性来表达词相似性[19]。深度学习模型在自然语言处理方面也得到很多的应用，最广泛的应用是利用神经概率语言模型来学习语言的表达，这其中就会产生词的表达，利用词的表达来分析词语之间的关系，然后进行短文本分析。

中英文在形式和结构上存在着较大的差别，在中文处理领域，主流的文本表示模型依然是向量空间模型，短文本在向量空间模型当中的处理主要是特征扩展，这种对于短文本的处理存在着语义扩大、维度灾难等缺点。

## 论文主要研究内容

本文在短文本分析当中关注电子商务用户评论短文本信息，因为由于电子商务是买卖关系，买家在进行评论时往往对购买的商品带着真实的情感进行商品属性描述，这是一种天然的进行商品属性分析和买家情感分析的绝佳短文本。网购当中，影响一次网购的因素有很多，例如品牌、网店信誉、网店评分等，而这些因素的真实情况是隐含在用户评论当中的，但是用户评论很多，网民在网购的时候往往是根据网店的已有评分和处于前几页的评论来得出对于一个商品的心理评价，这种评价是不全面的。对于电商平台而言，假货和负面信息较多的店铺是影响一个电商品牌信誉度的，对于电商卖家而言,不好的评论进行传播会带来品牌危机。如何利用自动化的手段，对于这些评论进行分析，一是给予普通消费者一个全面的商品评价，二是给予电商平台和卖家一个评价店铺和商品的方法，是值得研究的。因此，需要探索行之有效的方法，建立合理模型用来分析用户评论。

针对已有的短文本分析的方法主要聚焦于特征扩展而带来的信息湮问题，同时针对从评论短文本挖掘用户真实情感和意图的角度出发，本文在总结已有的研究成果的基础上，从针对情感分析的短文本表示的模型化出发，主要的研究内容如下：

1. 本文将通过研究各种文本表示模型的特点，分析各种模型的优缺点，针对电子商务评论短文本进行情感分析而提出更加合适的文本表示模型。
2. 本文将研究通过词相似度的方法进行文本模型降维处理的方法，本文在此基础上研究基于HowNet词相似度计算方法并加以改进，同时研究在语言模型下词的表示，进而通过这种词的表示去度量词相似度。通过实验验证这两种词相似度计算方法的在相同语境场景中的准确性和在不同词性中的准确性。
3. 为了更加准确的表示短文本，本文将研究文本表示模型当中的每个维度拥有自己的权值，权值与主题相关，本文将通过研究特征选择方法，给出维度的权值给定方法；同时不同的主题在空间当中具有不同的维度选择，给出主题相关的子空间维度选择公式。短文本的每个维度的值通过句子依存关系所得到的维度与相关附属对象的联系进而进行计算，对象的值代表了对象的情感倾向。
4. 本文通过抓取亚马逊商城的评论和评分信息，将验证本文提出的模型是否有效的表达评分信息和情感倾向，从而验证本文提出的模型的有效性。
5. 面向天猫商城用户评论的分析，利用本文提出的文本表示模型，分析其评论文本的情感倾向和关注热点。

## 本文的组织结构

针对应用于商品评论情感分析的词相似度的研究以及适合进行情感分析的商品评论短文本模型化表示研究，本文的主要组织结构安排如下。

1. 对本文的研究背景与意义进行了概述，同时概述了本文所做的主要研究内容和论文的组织安排。
2. 对本文涉及的一些基础技术的理论进行了相对全面的介绍，包括文本的表示理论、特征选择方法、文本分类方法，最后重点介绍了本文要用到的HowNet和神经概率语言模型。
3. 和第四章是本文的核心内容。第三章详细介绍了基于HowNet的知识结构而提出的改进的词相似度计算方法并进行验证。第四章在CBOW模型的基础上进行改进，并提出基于模型产生的词向量进行词相似度计算的各种方法，然后进行验证和比较。
4. 利用第三章和第四章的词相似度计算方法建立文本对象空间模型，并在此模型下提出了对象权值的计算、对象值的计算和对象维度的选择。然后通过实验进行数据验证与分析。
5. 面向天猫商城商品评论的特性，应用本文提出的文本对象空间模型设计和实现了一个对评论短文本进行分析的应用，并给出典型的功能的分析结果。
6. 对本文的研究进行了总结，提出了研究当中还可以改进的地方作为下一步工作的展望。

## 本章小结

本章对本文的研究背景与意义做了概述，同时总结了目前短文本分析的现状。最后对本文所做的主要研究内容进行了概括性介绍，并且说明了本文的论文结构安排。

# 文本分析方法与技术基础

本章将介绍短文本分析当中涉及到的一些方法和理论，包括文本如何表示，常用的特征选择以及中文文本分类领域常用的一些分类方法。知网和词分布式表示对于本文提出的短文本模型化表示具有基础性的作用，将对其进行重点介绍。

## 文本表示模型

要对短文本进行分析，首先要对文本进行数学建模，然后才能定量的研究文本。文本表示模型就是对文本做结构化的处理，使得文本可以用一些特征进行表示，即以一定的特征项来表示一段文本。

在实际的文本分类研究中，文本表示有多种方法，良好的文本表示方法能够更简洁、更准确的反映文本的内容，从而提高文本分析的效果。

1）布尔模型。布尔模型是一种二元简单模型，是一种类向量的模型。文档当中的特征的权重只有两个值，即，表示特征在文档中存在或者不存在。因为布尔模型下的特征是只有两个取值，所以无法利用其计算两个文档之间的更深层次的关系。

2）向量空间模型。一组文档可以用同一组维度去表示，这些维度形成了一个空间，这种表示方法被称为向量空间模型。主要思想是将一篇文档表示为一个空间中的一个点。文档通过向量空间模型进行表示之后，就可以方便的采用向量的有关计算方法对文档之间的关系进行分析。

在向量空间模型当中，常用的维度是字和词，一般都使用词语表示空间的维度，这种方式简单易用。对于每个文档，可以表示为如下形式：

 （2-1）

式中：表示的是文档中特征的权重。

采用向量空间模型的文本表示法最大的优点可以利用各种基于统计的方法对文档进行分析，但是由于维度的大小就是词典的大小，在词典中包含各种近似的词语，空间不能正交化且维数太高。

3）词汇链文本表示模型。一个文档当中，词语并不是随机的组合在一起的，而是围绕着一个主题或者特定的描述组织在一起的。词汇之间具有衔接关系，是文本的表层特性，主要通过文本单元之间的相关性来表现，相关性包括词汇的复现现象和搭配现象。

词汇链指的是文本中一系列的词共同组成的一个词汇序列，词汇序列表达了词汇之间的某种关系。词汇链模型将一个文本表示为几个由多个词组成词汇链，每个词汇链中的词由词汇链接关系联系在一起[20]。比如文本“小王感觉非常的累，于是他躺在一颗树下休息，有一片树叶落在了他的身上，紧接着一片又一片的树叶落下，最终他被一堆黄色的、绿色的的树叶包裹了起来”，该文本包含两条词汇链{累，躺，休息，包裹}和{树，树叶，落，黄色，绿色}，从这两条词汇链就可以分析出句子中各个成分之间的关系。常见的词汇衔接关系如表2-1[20]所示：

表2-1 词汇衔接关系类型

|  |  |
| --- | --- |
| 类型 | 定义 |
| 简单重复 | 词的简单重复 |
| 简单释义 | 一个词可以替换另外一个词，意思并未发生改变 |
| 复杂释义 | 通过两次词的关系去推断第三个词的关系 |
| 复杂重复 | 词语看起来相同，但是词性不同 |
| 语义关联 | 上下文、形容、共指 |
| 非词重复 | 人称和指示 |

词汇链模型构建中的关键问题是文本的句子结构分析、消除歧义以及合并，并且消除歧义和合并的时机不同会影响最终词汇链构建的准确度。

## 特征选择方法

无论使用哪一种文本模型表示法，由于字和词的数量往往很大，如果对文本特征项不加以选择，那么特征项的数量和词典的大小成线性关系，将造成维数灾难，并且在表征单个文本时，会产生特征非常稀疏的问题，因此特征选择是一个非常重要的环节。

1）TF\_IDF。TF（Term Frequency）是指词频，即某个词在特定的文档中出现的频率。如果只统计某个文档的词频，通过设定词频的阈值来找出此文档的关键词，但是结果得到的是是一些常用的词语，比如“的”。计算公式如下：

 （2-2）

式中：某个词j在文档当中的词频是词j在当中的出现次数除以文档当中所有词出现次数的和。

DF（Document Frequency）是指文档频率，即包含某个给定的词的文档的数量。通过设定阈值可以找到文档集的关键词或者相关词，但是通常会得到一些信息量不大的词会被保留，比如“我”、“是”、“的”等。所以采用另外一个方法来衡量——逆文档频率（Inverse Document Frequency，IDF），其计算公式如下：

 （2-3）

式中：表示文档总数；表示包含词语的文档总数。

两种方法结合起来，就既可以找到某一个文档中关键的词，同时它对于所有的文档都是关键的。

 （2-4）

TF\_IDF值高于一定阈值的词语可以作为特征被选择出来，认为它是具有更大权重的词语。

2）信息增益。信息增益（Information Gain，IG）的方法来自于信息论，认为不同的词语给系统带来的信息量是不同的，计算公式如下：

 （2-5）

式中：表示词语出现的概率；表示有词语的文档属于类别的概率；表示文档不有词语的概率；表示没有词语的文档属于类别的概率。

当一个特征带给整个系统的信息量比较大时，就会相对比较容易的采用信息增益的方式选择出来。

3）互信息。互信息（Mutual Information，MI）是一种关于信息量的度量，其意义是指词语在类别之间出现频率的统计。其近似计算公式如下：

 （2-6）

式中：A是指词语和类别为同时出现的文档数量；B是指词语没有出现而类别出现的文档数量；C是指包含词语出现但类别没有出现的文档数量数量；N是文档总数。

互信息对于低频词具有倾向性，低频词的MI值往往比较高，并且互信息的计算当中没有考虑TF信息，也会在一定程度上得到很多MI值相同的词语。

## 常用分类方法

分类的任务是寻找一个合理的分类器，能够根据输入的文本输出它相应的类别，因此一般的分类包含两部分：分类器的训练和分类器的测试。

经典的有监督分类方法有决策树、支持向量机、KNN、神经网络、朴素贝叶斯、随机森林等。这里介绍几种常用的方法。

### 支持向量机

支持向量机（Support Vector Machine，SVM）通过寻找最优的一个分隔超平面将不同类别的样本自动分开。最优超平面是通过最大化不同类别的样本当中距离最近的样本的间隔来确定的，这个间隔越大，分类器的泛化误差越小[21]。通过图2-1，来理解SVM。

图2-1中左图中的实线对于不同类别样本的分隔就比右图的效果好。最基本的SVM解决最经典的二分类问题。

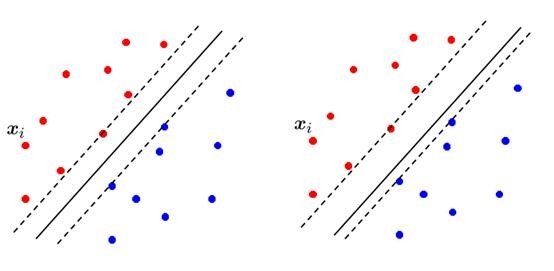


图2-1 SVM的最优超平面示意图

给定训练集，分类的超平面可以表示为，并且满足：，可以对分类间隔进行计算，在这个间隔的边缘超平面上的训练元组被称为支持向量（support vector），也就是说分类的超平面两侧的支持向量距离超平面一样近。目标是最大化，即是最小化，最终的问题转化为：

max ， （2-7）

解得支持向量的最优解，和权值向量以及偏置向量的最优解。最终得到的分类超平面为，且分类函数为：

 （2-8）

如果，则属于该类，否则不属于该类。当数据集线性不可分时，即将式2-8中的用一个函数来代替，称之为核函数，用来将输入映射到更加多维的空间，并在那个空间中求解。

### 决策树

决策树是一种表示决策转向和数据分流的结构，其中每一个非叶节点都表示对一个特征的选择性分流，每个分支表示对这个节点的测试的一个输出，叶子节点表示一个类[21]。一个典型的决策树如图2-2所示。

决策树分类算法的关键在于非叶节点的选择，即确定哪些特征作为树的某层的测试特征。常用的决策树分类算法有ID3、C4.5和CART。设需要在数据集D中选择测试属性，其期望信息是，其中是属于类i的非零概率，一般用测试集中出现的频率来估算。

1）采用信息增益来选择测试特征:

 （2-9）

式中：表示元组个数。使用信息增益最大的特征T作为测试特征。



图2-2 决策树示意图

2）采用信息增益率来选择测试特征：

 （2-10）

 （2-11）

采用信息增益率最大的特征作为测试特征。

3）采用基尼指数的方式来选择测试特征，基尼指数是指数据集的不纯度，定义为，其中是属于类i的非零概率。将某个特征中的值划分为两个部分，设一种划分的标识为j，划分后的数据集为，则这种划分的基尼指数为：

 （2-12）

### KNN

KNN是指K最近邻法，即对待分类的样本和空间上已有的所有样本进行相似度计算，然后在其中找出相似度最大的K个样本作为待分类样本的近邻，这K个近邻当中的多数属于哪个类别，就把待分类样本指定为这个类别[21]。

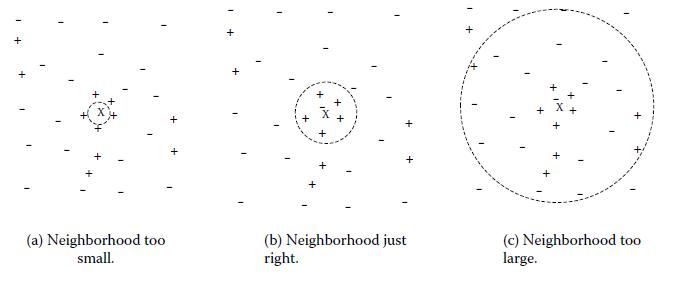


图2-3 KNN方法中不同K值对于分类的影响示意图

KNN方法在确定分类的决策时只依据最邻近的一个或者多个样本的类别来决定未知类别样本的所属类别，只与少量的相邻样本相关，对于类别区域交叉或者重叠较多的待分类样本集来说，KNN方法较为适合。KNN方法还可以用于回归，通过近邻样本的属性找到中心样本的属性特点，同时还可以根据样本之间的相似度来赋予权值。一般的采用点距离作为样本之间相似度度量的标准。KNN的缺点在于K值的选择，不同的K值会带来完全不同的结果。

### 神经网络分类器

神经网络是指用一系列的具有简单计算能力的单元（模拟神经元）来组成具有一定功能的结构模型。这些单元之间的连接可以随意定义，每个单元的自我调节（更新型学习算法）可以随意定义。其中常用的自我调节函数是阈值函数和Sigmoid函数[21]。

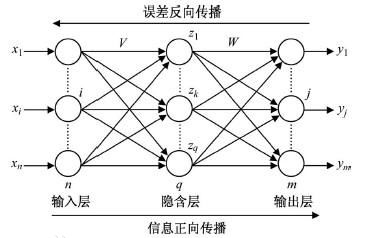


图2-4 3层前向反馈神经网络示意图

神经网络是自适应的、非线性的、非凸性的。输入层的单元接受外部的信号和数据，隐藏单元负责信息的处理和传递，信息的处理主要在于隐藏层的单元如何构建。

采用神经网络进行分类的思想是利用神经元之间信息的传递与反馈，使用已知类别的样本集合对各个神经元的参数通过训练进行更新，从而使得最终形成的神经网络能够判定输入的数据是什么类别。神经网络分类器的优点是构建灵活，可以自由的设定神经元的数量、神经元的连接方式等，可以很好的解决非线性可分的问题，缺点是参数的调节、初始值和学习步长的设定在理论上还不够完善，需要人工来不断的尝试，另外还存在着“过拟合”的问题。其中常用的用来分类的神经网路有多层感知机和BP网络。

## 知网（HowNet）

知网（HowNet）是一个研究语义的网络结构，在其中，用不可分的义原利用义项表达式来表示概念，用不同词性的概念来表示词语。利用义原之间、概念之间的关系来揭示词语之间的关系。

HowNet用一系列的义原来对每一个概念进行描述。HowNet的义原可以分为以下几个大类，如表2-2所示。

表2-2 HowNet义原类型

|  |  |
| --- | --- |
| 义原名称 | 义原含义 |
| Event | 事件 |
| entity | 实体 |
| attribute | 属性 |
| aValue | 属性值 |
| quantity | 数量 |
| qValue | 数量值 |
| SecondaryFeature | 次要特征 |
| syntax | 语法 |
| EventRole | 动态角色 |
| EventFeatures | 动态属性 |

HowNet的词描述举例如表2-3所示。

表2-3 HowNet词描述语言实例

|  |  |
| --- | --- |
| 词 | 描述 |
| 毛衣 | clothing|衣物,#body|身 |
| 材料 | material|材料,generic|统称 |
| 椅子 | furniture|家具,space|空间,@sit|坐蹲 |
| 屏幕 | part|部件,%machine|机器,\*look|看,#computer|电脑 |
| 吃饭 | eat|吃,patient=edible|食物 |
| 开车 | drive|驾驭,patient=LandVehicle|车 |

## 统计语言模型

语言模型是指自然语言只不过是某些特定的字母和词汇的有规律的组合，所以语言可以被一种数学模型所表达。

统计语言模型是用来计算一个句子是否是正常的自然语言的概率，通常基于一个语料库来构建。假设由N个词按按顺序构成一个句子，则同时出现的概率就是这个句子出现的可能性。利用贝叶斯公式，这个式子可以被链式的分解为式2-13。

 （2-13）

如果式2-13中所有的条件概率都可以算出来，那么给定的句子的概率就可算出，所以这些条件被称为统计语言模型的参数。

### n-gram模型

做一个n-1阶的Markov假设，假定一个词出现的概率只与它一定范围内的词相关，而这个数量被设定为n-1。

 （2-14）

当做了一个n-1的Markov假设之后，式2-13就发生了改变。

 （2-15）

这样的假定，使得单个参数的计算变得更为的容易，同时也使得参数总量变得更少。理论上n的值越大越好，但是n值越大，模型参数的增长呈指数级增长，假定词典的大小为NUM=200000，这个数字大约等于汉语的词汇量，模型参数的数量是NUM的指数函数（），显然n不能取的太大，在实际应用中，最多采用的是n=3。

采用n-gram模型可能会出现或者的情况，要采用专门的平滑化处理方式处理这种特殊情况。

### 神经网络模型

对于统计语言模型而言，可以考虑对问题进行建模，然后为其构造一个目标函数，然后对这个目标函数进行优化，从而求得一组最优的参数，最后将这一组最优的参数作为语言模型的参数，词向量是随着模型的建立而产生的。词向量有一种非常直观的表达方式，即将词表当中所有的词的总数n作为词向量的维度，如果某个词在词表当中的顺序是第j个，那么它的词向量的第j个维度设为“1”，其余位置设为“0”。One\_Hot Representation的表示下，词与词之间是孤立的，即便是同义词之间也看不到任何的关系，并不包含任何的语义信息，基于神经网络的Distributional Representation的目标是将词表示为一个n维的向量，所有的向量构成一个词向量空间，而每一个向量视为该空间的一个点，在这个n维空间上引入“距离”的概念，就可以分析词与词之间的关系了。

条件概率是语言模型参数的一个函数，神经网络就是一个良好的能够逼近这个函数的方法。Bengio等人提出一种神经概率语言模型，该模型当中用到了一个非常重要的工具——词向量[23]，它包括4个层。其中和分别是投影层与隐藏层之间以及隐藏层和输出层之间的权值矩阵，和分别为隐藏层和输出层上的偏置向量。其中是语料当中的一个词，是它的上下文。



图2-5 Bengio的神经网络示意图

借用n-gram的思想，将取为前面的n-1个词,这样就组成一个神经网络的训练样本。在输入层，将和都指定为一个m维的向量，具体的指定值可以随机，可以采用 one\_hot的方式，也可以用某种策略来确定；投影层的输出是一个维度为（n-1）m的向量，因为它是由输入层的n-1个m维向量按照顺序首尾拼接而成；用tanh函数做隐藏层的激活函数，即；输出层为。最终的输出做一个softmax归一化，就可以表示和的条件概率了。

 （2-16）

那么这个模型的待确定参数就包括各个词向量、、、、。这些参数的确定都通过神经网络的不断训练去优化目标函数所得，一般对最大似然对数函数进行优化。基于词向量的神经概率模型自带平滑化处理，式2-16的值是的。整个模型的计算集中在隐藏层和输出层之间的矩阵向量运算以及输出层的softmax归一化运算上，因此后续的相关研究工作中，基本上就是对这些部分进行优化的。

### word2vec中CBOW模型

Mikolov利用循环神经网络减少了前述的Bengio提出的神经概率语言模型中的参数，提出了RNNLM（Recurrent Neural Network Language Model，循环神经网络语言模型）[24]充分的利用词语产生的语境和场景信息来进行词语的预测，在预测当中形成语言模型，同时做词向量的更新。

word2vec就是Mikolov和自己的团队基于RNNLM在Google于2013年推出的一个词向量工具包，其中包含两大模型CBOW（Continuous Bag-of-Words Model）和Skip-gram（Continuous Skip-gram Model）[25]。

CBOW是利用当已经知道词w(t)的上下文w(t-2),w(t-1),w(t+1),w(t+2)的前提下预测w(t)；Skip-gram正好相反。本文重点介绍CBOW模型。

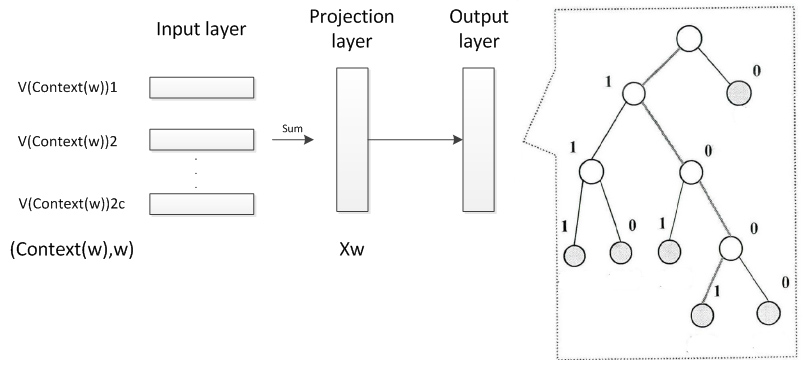


图2-6 CBOW模型的网络结构

输入层：包含2c个属于的词的词向量v，每个词向量的维度是m；投影层是将输入层的2c个向量做累加求和，输出层对应一颗二叉树，是以各个词在语料库中出现的次数作为权值构造的哈夫曼树，叶子节点有|NUM|个，对应词典当中的每个词，非叶节点有|NUM|-1个。

## 本章小结

这一章主要是针对文本分析对一些方法和理论做了一个概要的介绍。首先介绍了常用的文本表示模型，包括布尔模型、向量空间模型和词汇链模型；然后介绍了几种常用的特征选择方法，如TF\_IDF、信息增益和互信息；然后介绍了常用的文本分类方法，比如SVM、决策树、KNN和神经网络。最后重点介绍了HowNet和神经概率语言模型，为后续章节做准备。

# 基于HowNet的词相似度计算方法

本章结合传统的文本向量空间表示法、布尔表示法以及词汇链表示法，提出了适合对电子商务用户评论短文本进行文本分类和情感分类的文本对象空间表示形式，然后基于HowNet研究词相似度的计算用作空间降维。

## 面向商品评论情感分析的文本表示

中文文本分类问题研究常把单个词作为特征，往往会造成维数灾难。因为单个词作为特征用来表示一个维度，会导致空间并不是正交的，词与词之间存在着各种关系，尤其是近似关系，所以对于词语必须予以归纳，进行空间的降维处理。

本文分析的是电子商务用户评论短文本，这种特定场景下的短文本值得关注的是它所描述的事物以及对于这些事物的态度。布尔模型的优点在于可以直观的表示某一种特征是否存在，词汇链模型可以表示特征之间的关系，综合向量空间模型的优点，本文提出如式3-1所示的文本对象空间模型。

 （3-1）

式中：文本集合为；对象作为空间模型的特征维度，文本描述的对象的集合为；表示第i个文档在第j个对象维度上的情感值。关于这个情感值的计算将在第5章详细介绍。同时每个维度并不是等权的，权值与短文本的主题相关。

对象是指事物特征的概括和抽象的一个命名，相比于单个的词语，对象具有更高的抽象性。一个对象由一个或者一组词语描述。电子商务用户评论短文本是文本的撰写者通过对一个具体事物的特征的描述和评价来组成的。基于这样的一个客观事实，本文将从文本集当中抽取合适的词来表征事物的特征，形成事物的特征对象，这样解决了单纯以词语作为特征项时带来的维数过多、存在同义维度等问题。

## 基于HowNet的词相似度计算

本文的第二章已经对HowNet知识系统做了简单的介绍，在HowNet中，词的语义被概念所描述，概念被分解为各个义原。义原之间的关系最主要的关系是义原间由于语义上的概括和具象之间的区别而构成的上下位关系，这种关系可以将每个义原看做是一个节点，根据上下位关系，义原形成了一个有层次的树形结构。通过这个树形结构就可以对义原之间的联系进行分析，进而就可以分析概念和词。

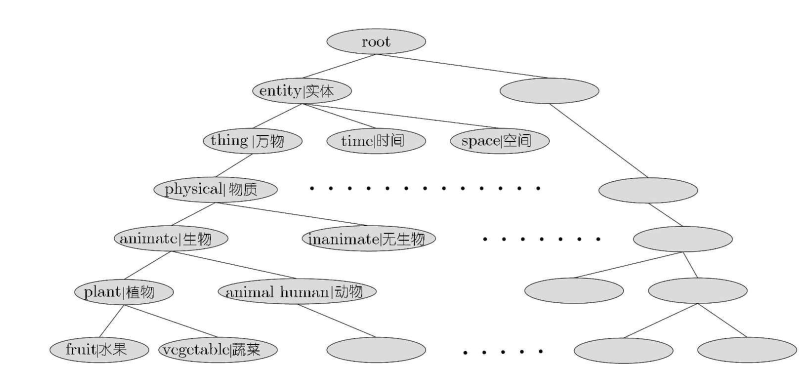


图3-1 义原层次树

### 基于义原距离和划分的词相似度计算

1）义原相似度计算：义原的上下位关系使得义原构成了一颗层次树，那么两个义原的相似度可以由两个义原在层次树结构中的路径所带来的距离决定，如式3-2[26]。

 （3-2）

式中：表示两个义原；表示两个义原的相似度；表示两个义原的距离。其中有两种计算方式：一是两个义原上溯到公共父节点时各自所走的最短路径的长度的和，每上溯一个父节点，路径长度加1。二是两个义原上溯到公共父节点时各自所走的最短路径的长度的差的绝对值。第一种表征了两个义原相互的绝对距离，第二种表征了两个义原之间相对于公共父节点的相对距离。当两个义原不在一颗树当中时，则取一个较大的值（一般是18-20）。

上述对于义原的分析方法简单易懂，但是太过于绝对化，可以加一个参数对仅仅考虑距离因素的公式进行调节，如式3-3[26].

 （3-3）

式中：为可调节参数，一般取值1.6；的计算采取路径和的计算方法。

采用式3-3时，会有相似度等同的问题，即虽然是不同组的义原，但是相似度等同。考虑2组义原，其中一组的=1+5=6，另外一组的=2+4=6，则两组义原的相似度相同，但是一个良好的义原相似度表征方法应该能够区分这两组义原。所以考虑义原节点在层次树中的深度作为参数。如式3-4[27]。

 （3-4）

式中：和表示两个义原深度。

采用式3-4计算义原相似度时，会导致深度较大的义原对参与计算时导致相似度较大，故加入调节参数进行修正。如式3-5[28]。

 （3-5）

式中：b为可调节参数，一般为2.0；同时将式3-4中的除以/2使得义原距离继续是主要影响因子。

2）概念相似度计算：一个概念可以分解为若干个义原，形成一个义项表达式，义项表达式可以按照义原特性分成4组[26]。第一基本义原组：义项表达式第一个位置的基本义原；独立义原组：义项表达式中第一个位置除外的基本义原；关系义原组：具有关系描述符的义原；具有符号修饰的义原。各个义原分组之间不产生交集，这样分组之后，按照组别可以分别计算组别的相似度。

 （3-6）

式中，是两个概念，是指概念在4个义原分组当中第i组的相似度。

一个义原分组可以包含多个义原，为了表征不同概念的同类义原分组的相似度，可以采用最大值法，假定，概念的第i个义原分组当中包含K个义原，概念的第i个义原分组当中包含T个义原。

 （3-7）

式中：将同类分组当中的属于两个概念的义原两两计算相似度，选择相似度最大的一个值作为义原分组的相似度的值。

有了4个同类义原分组之间的相似计算方法之后，就可以进行概念相似度的计算，由于4个同类义原分组互不相交、各自独立，所以采用加权求和的方式进行概念相似度计算。

第一个义原分组中包含的义原在层次树中位置较高，能够代表的归纳性更强，一般赋予更高的权重，按照这样思想，4个义原分组，可以按照在义项表达式中的位置，越是靠左，权重越大[29]。

 （3-8）

式中：表示第i个义原分组的权重，且，同时。

这种固定权重指定法实现简单，但是往往结果比较僵硬，不够灵活，所以既然越是靠左的义原组权重大，不如就加大靠左的义原分组的影响力。

 （3-9）

式中：每次权重乘积之前，先将某个义原分组相似度和其左边的义原分组的相似度做连乘操作，以加大靠左的义原分组的影响力；的意义与式3-8相同。

3）词相似度计算：假定两个词，每个词都由一个或者更多概念来描述，，同理由n个概念组成。

可以借鉴式3-7关于概念的义原分组计算模式，将描述的概念两两计算相似度，选取最大值作为词的相似度[26]。

 （3-10）

式中：，。

式3-10的词相似度计算方法并没有考虑词性对于词相似度的影响，相似的词在相似的上下文当中可以进行替换，可以进行替换的词必然词性是相同的。于是可以考虑按照词性划分分别计算，而后取最大值[30]。

 （3-11）

式中：，，同时具有相同的概念性质，比如名词（N）、动词（V）、情感词（ADJ）等。

### 改进的词相似度计算

1）改进的义原相似度计算：式3-5关于义原之间的距离的计算单一化，没有考虑义原相对距离。同时在计算时，没有考虑义原本身具有的情感倾向。所以对式3-5进行修改。

 （3-12）

式中：由两个义原的绝对距离和相对距离各自乘以自身的权重共同决定，一般取0.8，取0.2。

在不考虑义原的情感倾向时，根据式3-5计算的义原相似度有时候会产生跟常识不相符的结果。如表3-1所示。

“美”与“丑”是带有一定情感的义原且互相反义，“仪态”是不带有情感的义原，常识下“美”和“丑”的相似度应该更小，所以当不考虑义原的情感倾向而进行义原相似度计算时，会产生错误的结果。

表3-1 采用文献28（式3-5）计算的义原相似度

|  |  |  |
| --- | --- | --- |
| 义原1 | 义原2 | 义原相似度 |
| 美 | 丑 | 0.444 |
| 美 | 仪态 | 0.348 |
| 丑 | 仪态 | 0.348 |

首先对于原本的义原层次树进行改进，根据情感倾向，对每个义原节点赋予-1、0、1的情感倾向值，分别代表正向情感、中性情感、反向情感。当两个义原上溯到公共父节点时，各自路径上的情感倾向值进行加和处理，如果值为正，则说明义原的情感倾向值为正向，如果值为负，则说明义原的情感是倾向于负向的，中性情感同理。两个义原的路径上的情感倾向值的和值之间的差值可以作为调节参数加入式3-5。

 （3-13）

式中：表示义原在上溯到公共父节点时所经过的路径中最短的那条上的所有节点情感倾向值的和，同理。

 （3-14）

式中：q即为式3-13的计算结果，采用式3-12的计算结果。

当两个义原的情感倾向完全相同时，式3-14退化为式3-5。

1. 改进的概念相似度计算：文献29提出的式3-8和式3-9都基于一个假设：越是靠左的义原分组中包含的义原由于在层次树当中所处的层数比较的高，就越能表示抽象化的事物。本文认为这样简单的假设是不完全正确的。

表3-2 义项表达式举例

|  |  |
| --- | --- |
| 概念 | 义项表达式 |
| 美丽 | aValue|属性值,prettiness|美丑,beautiful|美,desired|良 |
| 丑陋 | aValue|属性值,prettiness|美丑,ugly|丑,undesired|莠 |

如表3-2所示，在常识认知当中完全相反的两个概念“美丽”和“丑陋”在义项表达式当中前两个义项是完全相同的，也就是说在高层抽象意义上“美丽”和“丑陋”是完全相同的概念，如果按照式3-8和式3-9的假设，“美丽”和“丑陋”两个概念当中靠后的两个能够反映概念不同的义原在概念相似度的计算中会被高层抽象义原几乎完全淹没，从而使得两个常识认知下完全相反的概念的相似度过高。

越是抽象的高层义原，其概括性和归纳性越强，随之带来的是比如情感倾向、具象描述的弱化，本文认为应该首先赋予概念的义项表达式中每一个义原一个权值，指导思想是越是靠右的低层义原其权值越大，同时由于高层义原的概括性强，在其权值较小的情形下加入影响因子。

将概念按照义原特性独立原则分成4个义原分组，如，，假设的第i个分组分别具有K和T个义原。则，其中是指第1 个概念的第个义原分组的第个义原和其权重，同理。

 （3-15）

式中：给出了第一个概念的各个义原的权值需要满足的条件，相同。

按照越是靠右的义原其权重越大的原则：

 （3-16）

式中：表示第个义原分组当中的义原个数，当i=1时，。

基于上述给出式3-16和义原分组，给出的计算公式。

 （3-17）

式中：分别表示两个概念的第i个义原分组当中的义原个数。

由于某个概念的第i个义原分组内部的义原权值和并不等于1，所以要进行归一化。

 （3-18）

式中：对于概念的第个义原分组当中的义原的权值进行了归一化处理，同理。于是式3-17变为如下形式。

 （3-19）

将每个概念的义项表达式进行义原特性分组，有时候会出现分组缺失的情况，比如分成了完整的4个组别，而只有3个组别。本文认为，当的义原组别i存在，而的义原组别i缺失时，的义原组别i的存在即是两个概念的差异化的存在，因为本文当中的相似度计算的值都归一化在[0,1]，则这种情况下的不对概念相似度的计算施加正向影响，则这种情况下。

假设和的义原分组i都缺失时，本文认为同时一个缺失的义原分组无法断定这两个义原分组的相似性，可能偏向相似，可能偏向不相似，则这种情况下，给出一个中性的结果。

计算最终的概念相似度时还是采取式3-9的形式，但是式3-9中只对义原分组i的权重给出了条件限定，没有给出取值方法。既然越是靠右的义原权值越大，则越是靠右的义原分组的权值越大。

 （3-20）

式中：某个义原分组i的权值是对两个概念各自的义原分组i的组内义原的权重加和进行求和，然后归一化除以2。

假设和的义原分组i都缺失时，本文认为，但是由于义原分组i的缺失，信息量贡献小，所以这个缺失的分组i的权值应该也小，不应该对最终的概念相似度计算结果施加过大的影响，则这种情况下，。

 （3-21）

式中：按照式3-20取值。

最终得到改进的式3-21，计算概念相似度时，对于靠右的义原和义原分组施加高权值，对于靠左的高层次抽象义原施加一定的影响因子。

3）改进的词相似度计算：式3-10没有考虑同一语境下的相似词之间的关系，式3-11充分考虑了词性的作用，但是同时忽略了如果两个词在不同的概念里拥有不同的词性恰恰是两个词的差异性所在。考虑本文最终的目的是为了对电子商务的用户评论短文本进行情感分类、文本分类和数据挖掘，语境是固定的，所以词性的作用非常重要，同时不同的词性对于相似度也要有影响。

（3-22）

式中：表示词性相同的概念计算，表示计算不同词性的概念相似度，是各自的权重系数。

由于语境比较固定，所以相同词性概念所得到的相似度对于词相似度的影响更大，只有当不存在相同词性概念时不同词性概念的影响力才会是1，所以且，当两个词不存在相同词性时：。

## 基于HowNet的词相似度计算结果实验分析

电子商务用户评论对一个事物进行描述的时候，必然是提出事物的某几项特征，然后有针对性的进行喜欢或者不喜欢的评价，所以在本文在建模时关注的是一个用户评论短文本当中的名词、情感词和程度词。

将分好词的文本集提取所有的名词、情感词和程度词，形成名词集合、情感词集合和程度词集合。每个词集合通过3.3所述的相似度计算方法，相似度大于一定阈值的词归于一个类别，用来表征同一个抽象对象。

为了验证在相似度计算当中的各种改进和优化的效果，选取一些具有代表性的义原、概念和词进行验证，同时利用HowNet的说明文件和同义词料库标注3200个词对进行分析，其中有2000个名词词对，1000个情感词词对，200个程度词词对。

### 义原相似度计算实验结果分析

选取一些典型的义原对计算相似度，与文献27、28提出的方法进行对比。

表3-3 部分义原相似度计算结果对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 义原A | 义原B | 义原相似度值 | | |
| 文献27 | 文献28 | 改进方法 |
| 1 | 锻练 | 活动 | 0.196 | 0.286 | 0.612 |
| 2 | 材质 | 材料 | 0.897 | 0.897 | 0.897 |
| 3 | 美 | 丑 | 0.524 | 0.444 | 0.090 |
| 4 | 喜悦 | 情绪 | 0.524 | 0.444 | 0.326 |

在义原相似度的计算中加入了情感倾向因子，在具有一定情感倾向的义原的相似度计算中表现良好，同时保持了中性情感的义原之间的相似度。

### 概念相似度计算实验结果分析

选取一些典型的概念对进行相似度计算，结果如表3-4所示。

表3-4 概念相似度计算结果对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 概念1 | 概念2 | 概念相似度值 | | |
| 文献29（式3-8） | 文献29（式3-9） | 改进方法 |
| 1 | 美丽 | 丑陋 | 0.783 | 0.815 | 0.104 |
| 2 | 高兴 | 害怕 | 0.344 | 0.286 | 0.081 |
| 3 | 恐惧 | 悲伤 | 0.444 | 0.444 | 0.556 |

在义原计算中采用加入情感因子，并且在概念相似度中对于不同的义原给予不同的权值，使得底层义原的影响力更高，高层义原分组在计算每个义原分组的相似度时对于结果施加影响，综合的结果使得概念相似度的计算更为精确。

### 词相似度计算实验结果分析

选取一些典型的词对进行相似度计算，结果如表3-5所示。

表3-5 词相似度计算结果对比

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 词语1 | 词语2 | 词相似度值 | | |
| 文献26 | 文献30 | 改进方法 |
| 1 | 计算机 | 电脑 | 1.000 | 1.000 | 1.000 |
| 2 | 材质 | 材料 | 0.897 | 0.897 | 0.897 |
| 3 | 男人 | 父亲 | 1.000 | 0.952 | 0.912 |
| 4 | 货 | 东西 | 0.278 | 0.278 | 0.463 |
| 5 | 喜欢 | 好 | 1.000 | 1.000 | 0.951 |
| 6 | 足球 | 运动 | 1.000 | 0.989 | 0.800 |
| 7 | 害怕 | 悲哀 | 0.856 | 0.856 | 0.444 |

改进后的词相似度计算公式（式3-22）的取值分别为0.9和0.1,保持了同义词之间的相似度，同时在分解的义原相似度的计算中反映了词的不同情感倾向，并且当词的概念属于不同的词性时，反映了其中的差异。

在标注的3200个词对中，词相似度计算准确率结果对比结果如图3-2所示，相似度阈值是0.5。与文献26、文献30的方法进行对比可以看出改进后的词相似度计算方法（式3-22）的效果相比未改进的其它两式有了很大的提升。但是准确率依然不是非常的理想，在进行实验的3200个词对当中，名词词对有2000个，本文的改进方法在对义原进行处理的时候在情感倾向因子的赋值时对于中性义原的情感倾向值为0，所以在实验集中名词词对占比比较大的情况下，可能会导致整体的正确率降低，所以将改进的词相似度计算方法在不同的词性的词对中进行对比，如图3-3。

图3-2 基于HowNet的词相似度计算准确率对比

改进前后的方法在不同的词性的词对中的结果：

图3-3 基于HowNet的词相似度准确率在不同词性中的对比

图3-3可以看出改进后的方法在情感词和程度词上的表现非常好，在名词的表现上效果相对较差，但是都比未改进的其它两式有了提升，尤其是在情感词和程度词的词相似度计算上，提升很大。因为很多名词具有多重概念，同时HowNet是一个字典形式的，所以对未收录的词无法进行处理，情感词和程度词在中文中所处的语境和应用相对稳定，名词集合是不断扩展的。

综上，本文所提出的改进后的词相似度计算，在计算词相似度进行分解的每一步都提升了相似度计算的准确率，最终反映在词相似度的计算上，名词相似度有了一定的提升，在情感词和程度词上的准确率提升很大。

## 本章小结

本章首先根据HowNet中义原根据上下位关系形成的层次树结构对原有的义原相似性度量方法进行了改进，然后根据概念的义原组成方式和义原层次树之间的关系改进了概念相似性的度量方法，最后在词相似性进行度量时考虑不同的词性贡献。实验表明本文改进后的相似度计算方法是有效的，可以作为构建文本对象空间时对象抽取与归纳的方法，但是由于HowNet是一个人为构建的知识网络，其中必然不能包含所有的词语，当遇到非收录词时就无法正常工作，下章将以机器学习作为方法研究词语的相似度计算。

# 基于词向量的词相似性度量方法

为了解决第三章的词相似度计算方法当中对于未收录词无法处理的情形，本章将主要介绍改进的CBOW模型和基于词的向量化表示研究多种计算方法，进行实验验证。第二章已经初步介绍了词向量的分布式表示和产生词向量的CBOW模型，当一个词可以被一个向量表征的时候，那么词与词之间的关系就可以用向量之间的关系来分析。一个能够产生良好的词向量的方法就成为了分析词关系的基础。

## 改进的CBOW模型

在CBOW模型当中，投影层是输入层的简单加和，考虑一个特定的词，其上下文中的不同词和之间的关系是不同的，所以不同的词对于的贡献也应该是不同的。如图4-1所示。

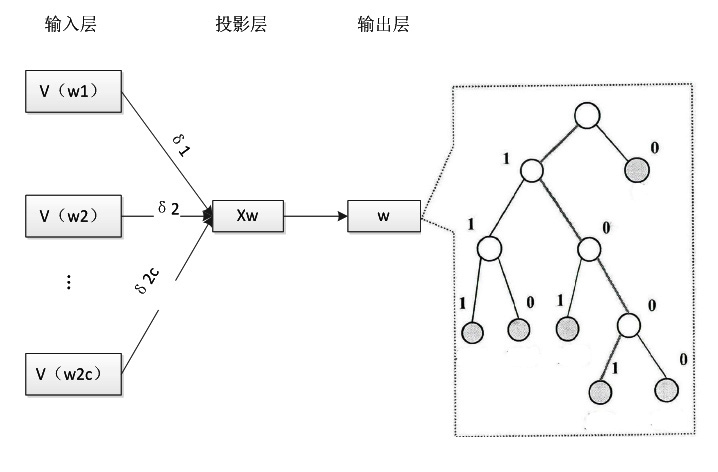


图4-1 改进的CBOW模型

基于此，则的计算变为加权累加。

 （4-1）

在正常的语言环境中可以做一个假设，即距离越近的词发生关系的可能性越大，且越近的词其关系越紧密，所以每个词的权值可以用词与词之间的间隔来度量，间隔越远的词对于中心词的贡献越小，基于词的间隔采用非线性计算权值的方式给出上下文中每个词的权值。，其中表示两个词之间间隔的词的数量。

对于语料库中的每个词，输出层即存在着一条从根节点开始到词对应的叶子节点结束的路径，而且这条路径是唯一的。路径存在着个分支，每个分支可以看做一次二分类，每一次的二分类就产生一个概率[25]，路径上所有概率的连乘就是。

 （4-2）

式中：表示路径上第个节点在哈夫曼树中由父节点分配的0或1构成的顺序码；表示路径上第个非叶节点的向量。

 （4-3）

式中：，用逻辑回归来表示当发生判断从而衍生出分支时的概率，作为二分类的概率。

设目标函数为，将式4-3带入。

 （4-4）



采用随机梯度法来优化目标函数[25]，设中除了双求和符号之外的部分为。随机梯度法的过程是：每次取一个样本，就对目标函数当中的所有相关参数做一次更新，这些参数包括向量、、、。

先考虑关于的梯度计算。

 （4-5）

于是的更新公式可以写为式4-6。

 （4-6）

式中：是学习率。

再考虑关于的梯度计算，由于和在中是对称的，所以它们的梯度公式类似。由于是input层所有的词的词向量的加和，那么更新每个词就要用到关于的梯度。

 （4-7）

与CBOW模型不同，改进后的CBOW模型的词向量的更新采用加权平均贡献。

 ， （4-8）

式中：表示的上下文的词个数。

改进后的CBOW模型的单样本参数更新算法为算法4.1。

算法4.1 改进后的CBOW模型下单样本参数更新算法。

输入：上下文

输出：更新非叶向量和词向量

Begin

；初始化；

；

for j=2 to 

；

；

；



for each 



End

## 词向量度量词相似性的方法

词向量使得是词可以看做在一个高维空间的点，相似的词在整体和单个维度上应该都是具有某种关系的，这种关系可以利用距离度量和相异系数度量，考虑不人工的预设阈值，可以利用分类的方法度量。

### 距离法度量词相似度

向量可以看做是高维空间的点，那么向量之间关系可以使用向量之间的距离去度量，距离越近越相似。设两个N维词向量和。

1）欧式距离。

 ， （4-9）

式中：表示词向量欧式距离，表示词相似度。

两个词向量的欧式距离越大，相似性越低，使用欧式距离的倒数来度量两个词向量之间的相似性。

2）余弦距离。

两个词向量在N维空间当中会形成的夹角，当这个夹角越小，则两个词向量越是接近，利用余弦值，余弦值越大的词对越近似。

 （4-10）

### 分类法度量词相似度

考虑将词映射为一个高维词向量之后，越是相似的词在相同维度上应该是相近的或者存在某种关系的，利用分类的方法去判断两个词是否相似时，则将词对看做分类器的输入，分类器可以将其判定为属于相似的类别或者不相似的类别。将词对的词向量进行处理，构造关系向量作为输入。

设两个N维词向量和。

1）差向量。考虑词向量单维度的差值，得到两个词向量的差向量，这个差向量当中值接近0的的维度越多，则两个词越接近。。

2）和向量。考虑假设当词向量的某一个维度是两个词向量的共同的重要特征，且这个特征经常出现在两个词的上下文当中，对两个词向量在按维度进行求和，相似的词在求和之后这样特征的趋势会加强。。

3）乘向量。在和向量的构造当中，认为两个词向量的共同重要特征可以度量两个词的相似性，考虑按维度对词向量求积，当结果为正的维度越多，且其值越大，则两个词在同样的特征上的趋势是相同的。。

4）连接向量。不预先设定两个词向量之间的相关程度或者某种关系，直接将两个词向量做连接，构成一个2N维度的向量，作为分类器的输入，让分类器去判别和寻找相关性。。

5）balAPinc度量[31]。如果两个词相似，则两个词的上下文存在交集，当这个交集的比例比较大，而且这个交集对于两个词而言是重要的维度交集，则两个词相似。。这个值作为分类器的输入，来判定词对相似度。其中表示的是两个词上下文交集的大小。

 （4-11）

式中：表示的是在词向量中维度的值，同理；当时，则，反之，同理。

表示上下文交集的重要程度。设为词向量中维度值大于0的个数，将一个词向量中的各位维度按照值降序排列得到，规定，即中第个维度的值等于。将归一到0-1的区间。。设另外一个词向量，表示的前个特征也在中，则，即如果一个维度的值在和中都大于0，则计入，否则不计入。

将归一化为，得到

 （4-12）

式中：若则，反之同理。

### 相异法度量词相似度

两个词向量的相同维度是表示同一个特征的，则在所有维度上对向量做出分解和选择，找到相异性，就得到了相似性。

设两个N维词向量和。

1）相对最大相异系数。设相对维度误差，这个值构成一个向量，然后对按照维度降序排列，取出其前个值，得到，对向量加权平均，得到最大相异系数。

 （4-13）

同理可以得到。当和的值比较小的时候，表明两个词相互之间的差异小，则两个词越是近似，得到相对最大相异系数。

2）奇异值分解。将两个词向量看做两个行向量构成一个矩阵，对这个矩阵进行如下分解：，其中是一个列正交矩阵，是一个对角矩阵，是一个列正交矩阵，这样的分解是为奇异值分解。将设为一个秩为2的对角矩阵，是因为两个词向量要么相似要么不相似，最大能够分成两类。得到的SVD分解之后，考虑，若或者，说明可以降维为矩阵，则说明两个词向量属于同一类。

## 基于词向量的词相似度计算实验结果分析

首先是词向量的训练，选择一种词相似度的度量方法，通过词相似度的计算来对比原始CBOW模型和改进的CBOW模型所产生的词向量在计算词相似度时的表现，进而对比4.2所述种度量方法的准确率，对比其优缺点，选择最优的方法作为最终的基于词向量进行词相似计算的方法。

### 词向量的训练

每一个词进行百度搜索，得到百度百科文档作为结果文档，假设没有对应的百度百科文档，就选取搜索结果的前5个非图片结果作为结果文档。对文档进行分词和清理，然后利用4.1所述的改进的CBOW模型训练词向量并保存。在每次训练新词时的输入初始化，先查询已有的词向量，若有，则直接使用，若没有，则随机初始化，当结果文档中出现词的所有上下文都已经被处理完成，则词向量训练结束。活动图如下。



图4-2 词向量训练活动图

### 基于词向量的词相似度计算结果分析

1）词向量训练算法比对。利用原始的CBOW模型、Skip-gram模型与4.1中的改进后的CBOW模型分别做训练相应的词向量。由于是采用词向量做相似度计算的，所以将词相似度计算的准确率作为评价标准，能够正确反映词相似度的即为更好的词向量。在度量词向量的优劣方面，经典的距离相似度计算方法简单，采用距离相似度来度量词向量的优劣，采用3.3所述的2000个名词词对、1000个情感词词对和200个程度词词对进行验证，训练100维的词向量进行计算。

图4-3 不同词向量的余弦相似度计算准确率对比

可见，4.1给出的加权的CBOW模型训练出来的词向量在相似度计算上有更好的准确率。

考虑词向量的维数大小对于词相似度的影响，训练50、100、150、200维的词向量，利用余弦相似度进行比较。结果如图4-4所示。

图4-4 不同维数的词向量的余弦相似度计算准确率对比

在词向量维数相同的情况下，改进后的CBOW模型都比原始模型所产生的词向量所计算的词相似度准确率高，在150维时效果最好，考虑到计算的时间性能，在实际的应用中采用效果比较近似的100维词向量。

那么不同模型产生的100维词向量在不同词性的词对的余弦相似度计算准确率上的比较如图4-5所示。

图4-5 不同模型词向量在不同词对上的相似度计算准确率对比

由图4-5可见，4.1给出模型在名词词对上的表现非常好，这是因为相似名词的上下文必然是相似的，而情感词和程度词由于在语言中使用灵活，上下文信息充满各种变化，所以难以用词向量的方式来计算相似度。

2）距离度量法。前述证明了4.1给出的模型会在相似度对比中得到更好的词向量，给出一些典型词语进行距离度量法的对比。

如表4-1所示，在词语相似度的表达上，余弦距离相似度要比欧式距离相似度的表达效果要好。

距离度量法的缺点在于需要人工指定相似度阈值，这个阈值的指定一般来说是0.5，但是并没有相应的理论支持。

表 4-1 距离法度量词相似度结果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 词语1 | 词语2 | 词相似度值 | |
| 欧式距离相似度 | 余弦距离相似度 |
| 1 | 计算机 | 电脑 | 1.000 | 1.000 |
| 2 | 材质 | 材料 | 0.762 | 0.853 |
| 3 | 女友 | 男友 | 0.426 | 0.673 |
| 4 | 货 | 东西 | 0.519 | 0.615 |
| 5 | 喜欢 | 好 | 0.736 | 0.864 |
| 6 | 足球 | 运动 | 0.854 | 0.873 |
| 7 | 害怕 | 悲哀 | 0.578 | 0.583 |

3）分类度量法。在HowNet中和同义语料库中抽取6000个词对，其中3000个相似词对，3000个非相似词对，1500个相似词对和1500个非相似词对构成训练集，其它的构成测试集。由于两个词只存在相似或者不相似的关系，采用天然的二分法分类器SVM，定义评价指标如下：定义混淆矩阵，其中表示当一个词对实际属于类别同时分类结果为的数量。其中类1表明是相似词对，类0表明不是相似词对。分别定义准确率、召回率和F值。，，，，，。然后进行权值综合。

，，，

， （4-14）

式中：。

由表4-2可以看出，两个词的关系向量能够反映两个词的相似度，其中和向量分类法表现最不好，差关系向量和balAPinc度量的表现最好，连接关系向量和乘关系向量的表现中规中矩。考虑对以上向量法进行组合。

表4-2 分类法度量词相似度的SVM分类器性能

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 向量 |  |  |  |  |  |  |  |  |  |
|  | 0.857 | 0.8 | 0.827 | 0.812 | 0.867 | 0.839 | 0.835 | 0.833 | 0.833 |
|  | 0.787 | 0.74 | 0.763 | 0.755 | 0.8 | 0.777 | 0.771 | 0.77 | 0.77 |
|  | 0.809 | 0.793 | 0.801 | 0.797 | 0.813 | 0.805 | 0.803 | 0.803 | 0.803 |
|  | 0.865 | 0.767 | 0.813 | 0.790 | 0.88 | 0.833 | 0.827 | 0.823 | 0.823 |
|  | 0.868 | 0.833 | 0.850 | 0.84 | 0.873 | 0.856 | 0.854 | 0.853 | 0.853 |

由表4-3可以看出，组合的关系向量进行分类法度量的效果比单个关系向量的效果更好，其中差向量和balAPinc度量的组合表现最好，延续其在单向量法中的良好表现。

表4-3 组合向量分类法度量词相似度的SVM分类器性能

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 向量 |  |  |  |  |  |  |  |  |  |
| + | 0.903 | 0.867 | 0.884 | 0.872 | 0.906 | 0.889 | 0.887 | 0.887 | 0.887 |
| + | 0.871 | 0.853 | 0.862 | 0.856 | 0.873 | 0.864 | 0.863 | 0.863 | 0.863 |
| + | 0.855 | 0.827 | 0.841 | 0.832 | 0.86 | 0.846 | 0.844 | 0.843 | 0.843 |

差向量可以反映两个词向量在每个维度上的差异，相似词的差异性在求差的过程中被消除，balAPinc得出两个词向量的最大重要交集，相似度的最大重要交集是比较接近全维的，所以这两种关系向量的组合向量性能非常好。

组合关系向量的SVM分类器性能用条形图直观的表示如下。

图4-6 组合向量分类法度量词相似度的分类器性能

4）相异法度量。选取表4-1所示的代表词语进行分析。

相对最大相异系数法点在于无法用一个确切的阈值去定义相似或者不相似，阈值指定过于困难。SVD方法的表现非常好，但是SVD的计算复杂，只有当词向量维度较低并且对于词差异性的精确度不敏感的环境中可以使用。

表 4-4 相异法度量词相似度结果对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 词语1 | 词语2 | 词相异性 | |
| 相对最大相异系数 | SVD分解后的矩阵E维度 |
| 1 | 计算机 | 电脑 | 0 | 1 |
| 2 | 材质 | 材料 | 2.652 | 1 |
| 3 | 女友 | 男友 | 9.567 | 1 |
| 4 | 货 | 东西 | 11.358 | 2 |
| 5 | 喜欢 | 好 | 4.812 | 1 |
| 6 | 足球 | 运动 | 2.015 | 1 |
| 7 | 害怕 | 悲哀 | 102.743 | 2 |

综上所述，通过改进的CBOW模型所训练的词向量能够更好的计算词相似度，在3种度量方法当中，分类法不需要人工指定相似度阈值，而距离度量法和最大相异系数法需要指定；分类法在计算难度上优于SVD分解，同时效果良好，所以分类法可以选择作为基于词向量进行词相似度计算的方法。

## 本章小结

本章对CBOW模型的投影层和词向量更新做了更改，基于词的向量表示，给出了多种基于向量的计算词相似度的方法。实验结果证明利用改进后的CBOW模型产生的词向量计算词相似度的方法是有效的，并且解决了第3章中的无法解决非收录词的问题。下章将根据第3章和本章所述，选择最优化的方法来构建文本对象空间模型。

# 基于词相似度的文本对象空间模型的建立和分析

本章将利用第三章和第四章提出的词相似度计算方法，在商品评论短文本的文本对象空间模型建立过程中选择最优的方法进行对象的抽取和归纳，之后给出每个对象维度的权值计算方法，最后在文本对象空间内对商品评论短文本进行分析。

## 数据集准备

本文研究的场景主要是电子商务平台中的用户评论数据，所以数据的来源选择为天猫商城和亚马逊商城。

1）数据采集：基于python+webdriver构建数据采集系统。主要采集的信息为：店铺名称、宝贝名称、宝贝总评分、评论文本。

由于天猫商城的用户评论是用户在下拉浏览、搜索、点击等操作中动态从服务器端访问生成的，所以必须采用webdriver来模拟人的各种浏览器操作来得到数据。其信息采集如图5-1所示：



图5-1 数据采集活动图

在关键字列表当中按顺序取出关键字，模拟输入和点击“搜索”，在商品的结果列表当中按顺序选择商品进行点击进入宝贝详情页面，在宝贝详情页面当中按顺序得到每一条评论并存储。

2）数据存储：采集的信息存储结构如表5-1和表5-2所示：

表5-1 商品信息表

|  |  |
| --- | --- |
| 字段 | 定义 |
| commodityId | 商品在数据库中的Id |
| shopName | 商品所在店铺名称 |
| commodityType | 商品所在类别 |
| commodityName | 商品名称 |

表5-2 评论信息表

|  |  |
| --- | --- |
| 字段 | 定义 |
| commentId | 评论Id |
| commodityId | 商品在数据库中的Id |
| commentInfo | 一条商品评论信息 |
| score | 评论的原始评分 |

本文通过上述方法收集了天猫商城11个品类280个商品共约28万条评论和亚马逊8个品类500个商品共约25000条评论。

表5‑3 天猫商城评论数据量

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 大衣 | 家居 | 健康 | 裤子 | 男鞋 | 美食 | 美妆 | 母婴 | 男装 | 女鞋 | 裙子 |
| 10062 | 43013 | 45730 | 21728 | 36869 | 22075 | 29477 | 18788 | 25532 | 15252 | 19376 |

表5‑4 亚马逊评论数据量

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 手机 | 家居 | 办公 | 美妆 | 食品 | 玩具 | 服饰 | 母婴 |
| 2982 | 3234 | 1541 | 4682 | 1249 | 2147 | 5894 | 3111 |

天猫商城数据用来做词集合的构建和最终的分析使用，亚马逊评论数据每个单条评论是带有评分的，所以用做训练集和测试集。

## 文本表示模型的建立

### 文本对象空间模型

对电子商务用户评论短文本建立文本对象空间模型时，采用名词对象作为空间的维度，采用情感词对象和程度词对象对维度的进行权值计算，每个名词对象由多个情感词对象来形容，每个情感词对象由多个程度词对象来加权，这样每种对象都对文本的表示有影响，模型如图5-2所示。



图5-2 商品评论短文本的文本对象空间模型

### 对象的抽取和归纳

将词和对象之间进行转化，需要进行对象的抽取和归纳。用算法5.1来说明对象抽取和归纳的过程。

算法5.1 对象抽取和归纳

输入：词集合（其中表示名词集合，同理分别是情感词集合和程度词集合）。

输出：对象集合（其中表示名词聚合而成的对象集合，其中每一个是由一个或者多个名词组成的聚合。同理分别表示情感词和程度词构成的对象集合）。

Begin

；

；

for each 

for each 

if  similar with one 



else

new 



End

以名词集合为例来说明名词对象集合的产生。选取一个名词，如果它和已经存在某个名词对象的相似度最大且大于阈值，则这个名词属于这个名词对象。如果不存在这样的一个名词对象，则新建一个名词对象，将这个名词加入[32]。不同的是，判断一个词是否属于一个对象时，基于HowNet的方法采用最小值法[33]，即当一个词和一个对象进行比较时，这个词和对象包含的所有词进行相似度计算，其中的最小值为词和对象的相似度；基于词向量的方法采用中心点法[34][35]，即当一个词和一个对象进行比较时，这个词和对象包含的所有词的词向量的平均值进行相似度计算。

整个对象抽取和归纳的活动图如图5-3所示。



图5-3 对象抽取和归纳活动图

采用HowNet的词相似度度量方法和采用词向量的词相似度度量方法在不同的词对中效果不同，下面将根据不同的词对选择不同的方法。

### 词相似度计算方法选择

1. 提出的基于HowNet的改进的词相似度计算方法和第四章提出的基于改进CBOW模型产生的词向量度量词相似度的方法各有优劣，在构建短文本的文本对象空间模型当中可以针对不同的情况选择不同的计算方法最优化模型的构建。

1）名词相似度计算方法选择。基于HowNet的词相似度计算方法在名词方面表现很好，但是存在着两大问题，一是未收录词无法进行计算，二是由于计算词相似度的时候是脱开语境的，虽然用多个概念可以去表征一个名词在不同语境下的表现，但是无法明确当前在比较相似度时的语境是什么，就会出现因为考虑多语境而导致相似度值变低的情况。基于CBOW模型产生的词向量，能够比较好的抽象表达一个词的上下文语境，同时可以解决未收录词的问题，所以在名词相似度计算上采用词向量的方法。根据4.3所述，各种基于词向量进行词相似度度量的方法各有优劣，距离法计算简单；分类法直观，并且可以同时计算多个词对；相异法另辟蹊径。由于相异法中的SVD分解计算复杂，本文采用的是100维词向量，所以不采用，同时相对最大相异系数法和距离法需要指定一个阈值去判定相似或者不相似，这个阈值的人工干预可能会导致相似度计算错误，所以最终采用分类法当中的向量组合法：。

2）情感词相似度计算方法选择。中文情感词是有限的，且意义非常明确带有明确的情感倾向，第三章提出的基于HowNet的改进的词相似度计算方法，对义原做了情感的倾向性处理，使得其在度量情感词的相似度方面表现非常好。对比之下，基于CBOW模型产生的词向量在情感词方面的相似度计算表现并不完全准确，因为同一个事物可以用完全相反的情感词去形容。如“我喜欢苹果”和“我讨厌苹果”，这两句话的针对“喜欢”和“讨厌”两个词的上下文完全一致，所以其产生的词向量也是相同的，这是不符合常识认知的。所以最终采用基于HowNet的改进的方法度量情感词的相似性。但是由于一些网络流行词语也可以去形容事物或者心情，这些词语并不包含在HowNet中，此时就要采用词向量的形式，首先人工指定一些词作为情感词标准词来表达各种情绪并利用4.3.1所述词向量训练方法得到新词和标准词的词向量，然后二者进行词向量相似度比较，新词近似于哪一个标准词就将这个新词替换为哪一个标准词。

3）程度词相似度计算方法选择。中文程度词相比情感词更加有限，同情感词一样，程度词存在着词不同而上下文语境完全相同的情况，所以词向量的方法失效，采用基于HowNet的方法去计算程度词相似度，同时由于程度词一般不会出现未收录词，所以也不必采用前面所述的对未收录词的处理方法。

### 对象值计算方法

为了能够正确的提取一条评论当中名词对象和情感词对象之间以及情感词对象和程度词对象之间的对应关系，要先分析评论当中各个成分之间的关系。

句子依存关系分析是为了分析句子的句法结构，目标是给定一个句子，然后给出一个带有句法标记的树形结构。句子的成分之间存在的关系被抽象为支配和被支配的关系。美国计算机语言学家罗宾逊提出了依存句法的4条公理[36]：首先，一个句子只有一个成分是独立的；其次，句子中的其它的成分直接从属于某一个成分；再次，任何一个成分都不能从属于两个或者两个以上的成分；最后，如果成分A直接从属于成分B，而C位于A和B之间，那么C直接从属于B或者AB之间的的某一成分，中心成分左右两面的其它成分互不发生联系。句法依存分析的结果是给出具有依存关系的各个词对，并给出表示各个词之间上下位关系的树结构。

斯坦福大学的StanforsNLP Group开发的开源NLP工具Standford Parser能够很好的完成句法依存分析的工作，同时还能给出词性标注。如“俄国希望伊朗没有制造核武器计划。”的句子依存分析结果为：

（ROOT（IP

（NP（NR 俄国））

（VP（VV 希望））

（IP

（NP（NR 伊朗））

（VP

（ADVP（AD 没有））

（VP（VV 制造））

（NP（NN 核武器） （NN 计划））））））

（PU 。）））

有了基于句子依存关系进行句法分析的结果词对和树结构，就能够对电子商务用户评论短文本当中出现的名词和情感词以及情感词和程度词的关系词对进行准确的提取。这些词对构成的树结构图5-2所示。

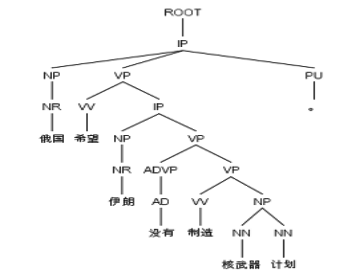


图5-4 Standford Parser句法分析举例

采用式3-1的表示，给出对象值的计算公式。

文本集合为；文本描述的对象的集合为，名词对象作为空间模型的特征维度；是第i个文档在第j个对象维度上的权值。

在图5-1中对于一个维度施加影响的是与之依存的情感词的抽象对象，对情感词施加影响的是与之依存的程度词抽象对象,利用链关系进行对象值计算[37][38]。

采用人工预定义小字典方式得到初始一部分情感词和程度词的值。的首先人工预定义正向情感情感词字典和反向情感情感词字典，这里的情感词字典是广义的情感词字典，包括了形容词和动词在内，很多动词是带有情感倾向的[39]，比如“喜欢”和“讨厌”。分别在其中预定义3大类的代表性词语，每个大类只有2到3个代表性词语。

表 5-5 预定义情感词字典

|  |  |  |  |
| --- | --- | --- | --- |
| 类别 | 程度 | 字典 | |
| 正向 | 反向 |
| 1 | 极其|extremely|1.0 | 爱，好，赞 | 悔、哭、骂 |
| 2 | 很|very|0.8 | 喜欢，满意，快 | 慢、讨厌 |
| 3 | 较|more|0.5 | 可以，行，一般 | 为难、纠结 |

其中程度中的数值表示程度权值，正向取正值，反向取负值，定义为。

其次，预定义正向和反向程度词程度字典：，其中预定义6大类的代表性词语，每个大类只有2到3个代表性词语。

再次预定义程度词字典，预定义7大类的代表性词语，每个大类只有2到3个代表性词语。每个类别有初始定义的值，定义为，其中1到3类取大于1的正数，4和5取小于1的正数，第6类取大于-1的负数，第7类取-1。

表 5-6 预定义程度词字典

|  |  |  |
| --- | --- | --- |
| 类别 | 程度 | 字典 |
| 1 | 极其|extremely | 极其、非常、最 |
| 2 | 很|very | 很、尤其、着实 |
| 3 | 较|more | 还、越来越 |
| 4 | 稍|little | 多多少少、稍微 |
| 5 | 欠|insufficiently | 相对、没怎么 |
| 6 | 超|over | 过分、超 |
| 7 | 否定|no | 不、没有 |

当根据句子依存关系，得到一个句子的名词对象为，对名词对象其施加影响的情感词对象为，设对其中一个情感词对象施加影响的程度词对象为。则对其中一个名词对象计算权值的公式为：

 （5-1）

将权值归一处理到[0,1]内，得到最终的对象值计算公式：

 （5-2）

电子商务评论网站的打分通常为1到5星，将其转化为[0,1]区间的值，为{0.2,0.4,0.6,0.8,1.0}，这样便于和本文提出的对象值进行比较。

### 主题相关的对象维度选择

当一条评论经过处理，成为文本对象空间的一个点时，空间的维度远多于这条评论评价的商品所涉及的维度，所以必须针对每一个商品确定它的品类，根据品类构成的主题词选择相关维度，实现降维。

5.1进行数据收集时，每条评论都有来自于哪个商品而这个商品属于什么品类的数据，这个品类数据就可以作为维度选择的主题。

将每个品类数据进行百度搜索，搜索出来的前10篇文档作为结果文档[40][41]。在结果文档当中采用2.2.1所述的TF\_IDF方法选择特征名词[42][43]，并采用5.2.3所述的词与对象进行相似度比较的方法，将这些特征名词映射到已有的名词对象维度上，则这些对象维度就是与这个主题相关的对象维度，与这个主题相关的所有商品评论都用这些对象维度来表征。

算法5.2 主题相关的维度选择

输入：一个主题Title

输出：主题相关的对象集合

Begin

；;

利用TF\_IDF方法得到Title的候选特征集合

其对应的TF\_IDF的值为TF\_IDF=

for each  in 

if  similar with 

push  into 

push  into 

最终得到和其对应的TF\_IDF的值集合

END

## 模型实验结果分析

1）整个文本对象空间生成的过程为：

（1）数据收集。

（2）中文分词。

（3）初始化词集合。

（4）名词集合采用改进的CBOW模型所产生的词向量利用分类度量法进行对象抽取和归纳。

（5）情感词和程度词采用基于HowNet的改进的词相似度计算方法进行对象抽取和归纳。

（6）得到对象集合，名词对象作为文本对象空间模型的维度。

2）然后给出评论短文本在模型当中的处理过程：

（1）评论短文本输入到文本对象空间模型。

（2）对每个文本主题进行主题相关的维度选择。

（3）分析每个评论文本的成分依存关系，得到对象链。

（4）利用对象链进行文本的对象值计算。

（5）对文本进行情感分析等。

对于上述建立的文本对象空间模型中的评论文本，要得到评论的情感倾向，采用两种分析方法，第一种是单条分析法，即测试集中每一条评论在这个模型下的评分是否和测试集已有的评分接近，然后看一组的总评分是否接近，单条分析法可以给出单条评论的具体评分；第二种是分类法[44]，将一组评论分成5个类别，每个类别按照1到5星给出评分，看是否和总评分接近，这样的方法更加贴近于实际应用当中的用户评分方法。

### 情感分析评价指标

1）单条分析从广义上来说也是一种分类，分为正确类和不正确类，由此可以采用分类的评价指标：准确率，只考虑正确类的比率。

 （5-3）

式中：为分为正确类的评论个数，为所有评论个数。

其中，判断一条评论是否被分为正确类，测试集数据中按照每个对象维度的值进行加权求和得到模型评分，每个对象维度所占的权值为进行主题维度选择时的每个对象维度的TF\_IDF系数的归一化。主题相关对象维度集合为，则每一个对象维度的权值为：

 （5-4）

利用公式5-2得到每个对象值，如果，则为正确类，即测试集中一条评论所有对象维度的加权和与本来应有的评分结果之间的差值在0.05以内，则表明采用模型的单条评分是正确的。

同样的，将评论按商品分组输入模型，组中评论的模型评分累加和的平均值与已有的评分进行对比，如果差值在0.05以内，则表明采用模型的组评分是正确的。

单条分析法可以采用表5-4所给的所有数据直接做模型测试。

2）分类法的评价指标采用经典的评价指标：准确率、召回率和F值。定义混淆矩阵，其中表示评论实际属于类别同时分类器结果为的数量。

 （5-5）

式中：被正确分为第i类的评论数和所有被分为第i类的评论数的商为第i类的准确率。

 （5-6）

式中：被正确分为第i类的评论数和所有应该被分为第i类的评论数的商为第i类的召回率。

 （5-7）

进行加权综合：

，，，（5-8）

良好的分类器，可以使得准确率、召回率和F值都尽可能的高。在进行分类器训练和测试之前，每个维度的权值和维度值进行乘积。

### 实验结果分析

1）单条分析法从单条评论评分的准确率和分组评论评分的准确率两方面验证模型。利用表5-4所有的数据进行对象的抽取和归纳从而形成文本对象空间，再通过主题相关的对象维度选择为每一个主题建立子空间，先将每条评论输入空间，得到每个对象维度上的对象值，然后在每个子空间里进行准确率验证，结果如表5-7所示。

由表5-7可见分组评分的准确率比较依赖于单条评分的准确率，这是因为分组评分只是单条评分的组内加和平均。手机、服饰、食品的单条评分准确率很高，而母婴、美妆、家居的单条评分准确率相对不高，这是由于主题设定的范围造成的影响，准确率高的主题往往是范围较为狭窄的主题，在较为狭窄的主题当中，利用TF\_IDF方法更能找到明确的与主题相关的属性，同时在对象空间里能够相对容易对应到已有的对象维度。

表 5-7 单条分析法准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 品类序号 | 品类 | 准确率 | |
| 单条评分 | 分组评分 |
| 1 | 手机 | 0.89 | 0.86 |
| 2 | 家居 | 0.63 | 0.59 |
| 3 | 办公 | 0.61 | 0.68 |
| 4 | 美妆 | 0.56 | 0.57 |
| 5 | 食品 | 0.83 | 0.87 |
| 6 | 玩具 | 0.72 | 0.69 |
| 7 | 服饰 | 0.91 | 0.89 |
| 8 | 母婴 | 0.58 | 0.61 |

2）分类法，利用前述已经建立好的文本对象空间模型和主题相关的子空间以及相应的每条的评论的对象值，在每个子空间里将评论分为5个类别，分别代表1星到5星，对应的标准评分对应为{0.2,0.4,0.6,0.8,1.0}。在表5-4中随机抽取8个品类共15000条评论进行分类器的训练，10000条评论进行测试。分类选择KNN、基于检验的CHAID决策树、多对多SVM三种。如表5-8所示。

由表5-8可以看出，KNN的分类效果最不好，SVM的效果中等，而CHAID的效果最好，同时与1）相似，手机、服饰、食品的相关分类器性能都属于比较好的，原因也是相同的，这3个品类的主题相对没有那么宽泛，特征明确。而因为对象维度是有权值的，KNN单纯从空间中点的距离因素来考量点的分类，放大了具有高权值的维度的影响因素；SVM通过多对多的分类器一定程度上规避了维度权值不同带来的影响，CHAID在分裂属性的选择上能够体现维度的权值。

表5-8 KNN\CHAID\SVM分类器性能

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 品类 | KNN | | | | | CHAID | | | | | SVM | | | | |
|  | |  | |  |  | |  |  | |  | |  | |  |
| 手机 | 0.814 | 0.764 | | 0.784 | | 0.891 | 0.884 | | | 0.885 | 0.853 | 0.815 | | 0.829 | |
| 家居 | 0.732 | 0.684 | | 0.701 | | 0.790 | 0.763 | | | 0.771 | 0.789 | 0.76 | | 0.769 | |
| 办公 | 0.733 | 0.662 | | 0.689 | | 0.786 | 0.733 | | | 0.753 | 0.776 | 0.719 | | 0.740 | |
| 美妆 | 0.689 | 0.603 | | 0.636 | | 0.745 | 0.678 | | | 0.703 | 0.702 | 0.623 | | 0.653 | |
| 食品 | 0.755 | 0.716 | | 0.729 | | 0.801 | 0.771 | | | 0.782 | 0.782 | 0.724 | | 0.746 | |
| 玩具 | 0.791 | 0.713 | | 0.732 | | 0.792 | 0.718 | | | 0.736 | 0.788 | 0.710 | | 0.729 | |
| 服饰 | 0.882 | 0.837 | | 0.854 | | 0.938 | 0.933 | | | 0.934 | 0.893 | 0.861 | | 0.873 | |
| 母婴 | 0.725 | 0.607 | | 0.635 | | 0.778 | 0.691 | | | 0.712 | 0.748 | 0.644 | | 0.670 | |

考虑到维度具有权值，选择朴素贝叶斯法和BP神经网络进行处理。结果如表5-9所示。

表5-9 朴素贝叶斯\BP网络分类器性能

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 品类 | 朴素贝叶斯 | | | BP网络 | | |
|  |  |  |  |  |  |
| 手机 | 0.887 | 0.886 | 0.881 | 0.893 | 0.891 | 0.892 |
| 家居 | 0.78 | 0.756 | 0.759 | 0.812 | 0.772 | 0.783 |
| 办公 | 0.79 | 0.74 | 0.759 | 0.793 | 0.745 | 0.761 |
| 美妆 | 0.758 | 0.705 | 0.726 | 0.816 | 0.817 | 0.816 |
| 食品 | 0.812 | 0.795 | 0.797 | 0.877 | 0.864 | 0.867 |
| 玩具 | 0.775 | 0.706 | 0.731 | 0.824 | 0.819 | 0.821 |
| 服饰 | 0.905 | 0.896 | 0.893 | 0.914 | 0.926 | 0.914 |
| 母婴 | 0.776 | 0.703 | 0.731 | 0.843 | 0.838 | 0.836 |

由表5-9所见，朴素贝叶斯法的效果与CHAID相当，而BP神经网络的方法表现最好，朴素贝叶斯用条件概率来体现了维度的权值，BP神经网络规避了单纯的空间距离因素导致的分类不正确。

分类法能够给出每一条评论属于1星到5星的哪一个类别，这种方法贴合于实际应用当中用户打分时的习惯，从评论文本的实际内容出发得到星评分，规避了打分者的错误点击或者受到第三方干扰下的星评分与实际评论不符合的情况。

下面根据平均的准确率、召回率、F值给出各个方法的比较，给直方图的形式呈现。

图5-5 各种分类器平均性能比较

综上所述，基于本文所给的文本对象空间模型，可以得到每条评论的情感评分，采用基于机器学习的分类方法，各种分类器的表现都比较好，其中CHAID决策树和BP神经网络的分类效果最好，都可以作为本文模型的分类器使用。

## 本章小结

本章首先根据第3章和第4章提出的词相似度计算方法，针对电子商务用户评论短文本的特点给出了相应的相似度计算方法的选择进行对象的抽取和归纳；进而根据句子依存关系给出了对象值的计算公式；然后建立了主题相关的对象维度选择方法；进而实验验证了利用模型对商品评论短文本情感打分的有效性。下章将根据天猫商城商品评论的特性分析，利用文本对象空间模型对天猫商城的商品评论进行分析。

# 文本对象空间模型在天猫商品评论分析中的应用

本章将分析天猫商城用户评论和评分的特点及其对普通用户产生的影响，分析天猫商城评论系统所带来的问题，进而利用B/S架构对前述章节的词相似度计算、文本对象空间模型建立、对象值计算、主题相关维度选择等过程进行实践，应用于天猫商城用户评论进行分析，从而解决天猫商城评论系统对普通用户的错误导向。

## 天猫商城商品评分系统的特点分析与问题

### 天猫商城商品评分系统的特点

天猫作为中国最大的电子商务平台，是阿里公司精心打造的精品电子商务平台，以各类商品的品牌性入驻作为特性，主打精品、正品的理念。平台上的众多商品都有用户的评论，几乎每一次电子商务买卖的达成都会产生一条或者数条关于商品的评论。

1. 天猫商城的商品评论涵盖多个品类，每个品类拥有多个商品，商品评论是一个数量很大的数据集，存在着挖掘和分析的价值。
2. 天猫商城的评论文本当中包含着每一个用户对于特定商品的喜好的描述，关注点的描述。
3. 当用户在天猫上进行评论和评分时，最终呈现给其它用户的信息是评论信息而没有评分信息。
4. 当用户在进行购物行为时，会不自觉的被商品和店铺的评分所影响，用户的购物决策不再仅仅由自身简单决定。
5. 天猫商城的评论和评分系统存在着能够对商品评分进行二次改动的功能，这个功能往往被卖家所利用，通过回电的方式请求某些打低分的买家进行改分，所以导致商品的评分存在着水分。
6. 天猫商城的卖家往往为了通过拉高评分来拉高销售额从而通过买水军的方式对旗下商品进行高评分，掩盖真实用户的真实评分。

### 天猫商城商品评分系统存在的问题

上述对于天猫商城评分系统的特性分析，可以得出天猫商城评分系统的如下问题：

1. 天猫商城的商品评分和店铺评分存在着水分，会误导消费者的购物决策。
2. 天猫商城呈现给普通消费者每一条评论都不带评分信息，在大量的评论文本面前，普通消费者往往找不到自己真实想要的信息。
3. 天猫商城的每一个商品的置顶评论往往是一些好评或者超级好评，大量真实的评论隐藏在多个评论页面当中，消费者往往看到的是卖家想要给消费者看到的评论，这些评论也会误导消费者的购物决策。
4. 天猫商城依托淘宝网建立，里面不可避免的充斥着假货，而这些假货商品可以通过买水军的方式拉高评分，从而拉高销售额，掩盖自身是假货的事实，而消费者往往在购买了商品之后，即便发现了是假货打了低分，也由于卖家回电要求改分的各种骚扰而不得不给一个假货偏高的分数。

## 天猫商品评论分析的功能需求

### 天猫商品评论分析的目标

通过6.1对于天猫商城评论系统的特性分析和其带来的问题分析，本章应用的主要目标是解决以下问题。

1. 通过对天猫商城的用户评论进行分析得到用户评论当中隐含的真实情感和意图，给每一条评论一个符合其真实意图的评分。为普通消费者提供一个天猫商城评论信息真实情感倾向和意图分析的应用。
2. 通过对天猫商城当中一些品类的用户评论数据进行分析，得到这些品类中消费者真实关注的方面和商品属性。为普通消费者提供一个决策是否购买一个商品的考虑的方向性指导。

### 需求描述

对天猫商品的用户评论进行建模和分析可以分为两大类的需求。

1）文本对象空间模型建立。即通过已有的电子商务用户评论建立文本对象空间模型，同时建立已有的主题和对象维度之间的选择关系。

2）在文本对象空间模型中分析未标注分数的文本。即通过已建立的文本对象空间模型，能够对单条、多条电子商务用户评论进行分析，典型的分析包括情感分析和关注点分析。

### 需求分析

1. 应用主要提供给数据分析人员使用，用例如图6-1所示。



图6-1 商品评论情感分析用例图

“文本预处理”就是从原始的评论文本当中通过分词和数据清洗得到词集合。用例描述见表6-1。

表6-1 文本预处理的用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 文本预处理 |
| 描述 | 通过原始的评论短文本建立词集合 |
| 参与者 | 数据分析人员 |
| 前置条件 | 原始评论短文本已存在 |
| 后置条件 | 确定名词集合、情感词集合、程度词集合 |
| 操作流程 | 1. 评论短文本进行分词 2. 分词结果去掉标点符号和去停用词 3. 建立名词集合、情感词集合、程度词集合 |

“建立文本对象空间模型”就是从数据库中的评论语句中通过词相似度的计算进行对象的抽取和归纳，以名词对象作为空间维度，通过主题相关维度选择计算方法，为数据库中每个已知的主题进行维度选择和维度权值分配。在这个用例当中主要使用了第三章和第四章中提出的词相似度计算方法和第五章提出的主题相关维度选择的计算方法。

建立文本对象空间模型是进行文本分析的基础，用例描述见表6-2。

表6-2 建立文本对象空间模型的用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 建立文本对象空间模型 |
| 描述 | 通过原始的词集合建立文本对象空间模型 |
| 参与者 | 数据分析人员 |
| 前置条件 | 原始评论短文本已存在 |
| 后置条件 | 确定文本对象空间模型的维度  确定与已存在的主题相关的维度 |
| 操作流程 | 1.词相似度计算  2.对象的抽取和归纳  3.使用主题相关维度选择计算进行主题相关的维度选择和权值赋予 |

“电子商务用户评论分析”是使用已经建立的文本对象空间模型和分类器对未标注的评论短文本进行情感分析和关注点分析。在这个用例当中主要使用了第五章提出的对象值计算方法和已经训练好的分类器，通过电子商务用户评论分析应用于天猫商城可以得到评分短文本所表达的真实信息。

电子商务用户评论分析的是利用第五章所述的情感打分分类器进行分析，同时加入了单维度的实值统计，也就是关注点的统计。用例描述见表6-3。

表6-3 电子商务用户评论分析的用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 电子商务用户评论分析 |
| 描述 | 通过分类器分析未标注的电子商务用户评论 |
| 参与者 | 数据分析人员 |
| 前置条件 | 1.文本对象空间模型已建立。2.分类器已训练完成 |
| 后置条件 | 商品用户评论分析完成 |
| 操作流程 | 1.利用已经训练好的评分分类器给出文本情感评分结果  2.统计计算评论文本的关注点 |

“对象值计算”使用已经建立好文本对象空间模型，为每一条评论短文本计算其在文本对象空间中的值，这个值可以使得一条评论短文本成为文本对象空间中的一个点。用例描述见表6-4。

表6-4 对象值计算的用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 对象值计算 |
| 描述 | 通过评论短文本的对象链得到对象值。 |
| 参与者 | 数据分析人员 |
| 前置条件 | 文本对象空间模型已建立 |
| 后置条件 | 确定评论短文本的对象值 |
| 操作流程 | 1.评论短文本进行依存句法分析  2.根据句法分析的结果文件和文本空间模型建立时归纳的对象得到短文本的对象链  3.使用预定义字典和对象链进行对象值的计算。 |

“分类器训练”是使用已经标注过的评论短文本进行分类器的训练，为后续的文本分析做好准备。在这个用例当中主要使用了第五章提出的对象值计算方法和效果良好的分类器选择。用例描述见表6-5。

表6-5 分类器训练的用例描述

|  |  |
| --- | --- |
| 描述项 | 说明 |
| 用例名称 | 分类器训练 |
| 描述 | 通过标注的用户评论短文本训练分类器 |
| 参与者 | 数据分析人员 |
| 前置条件 | 1. 文本对象空间模型已建立 2. 评论短文本已标注 3. 评论短文本的对象值已计算 |
| 后置条件 | 确定用于文本分析的各种分类器 |
| 操作流程 | 1.训练各种分类器 |

1. 上一小节通过用例分析描述了对天猫商城商品的评论进行分析的功能模型，下面将详细介绍进行评论短文本分析的行为模型。按照前一小节的对于功能的划分，给出关键功能的行为分析。

文本预处理阶段主要产生的是短文本的分词结果所形成的词集合。经过3个步骤：文本分词、数据清理、词集合产生。



图6-2 文本预处理活动图

文本对象空间模型的建立是利用词集合最终形成对象集合，并且为每一个主题建立子空间。经过主要的3个步骤：词相似度计算、根据词相似度结果进行对象的抽取和归纳、主题相关的名词对象维度选择。



图6-3 文本对象空间模型的建立活动图

对象值计算的目标是得到评论短文本的对象值，经过主要3个步骤：评论短文本经过依存句法分析得到词链，词链被映射成为对象链，对象链调用预定义情感字典得到对象值。



图6-4 对象值计算活动图

分类器训练的目标是根据一组已经计算好对象值的评论，得到进行情感评分的分类器，主要的步骤为利用已经设定好的分类器模型，通过输入一组评论作为训练集，更新分类器模型的参数，虽然分类器模型的参数的内部更新是一个复杂的过程，但是从外部来看，分类器行为的本身很简单，在此不做赘述。

电子商务用户评论分析的目标是利用分类器，将输入的一组评论进行情感打分，并且统计每个对象维度的实值个数和正负情感比例。主要经过2个步骤：评论输入分类器得到情感分类和类别对应的分数、统计一组评论的单维度的实值个数和正负情感比例，行为过程比较简单，在此不做赘述。

## 天猫商品评论分析应用的设计

### 天猫商品评论分析应用的框架结构

根据6.2所描述的对天猫商城商品评论进行情感分析的功能模型和行为模型，可以得到一个3层的框架结构。整个框架结构如图6-5所示。

图6-5 商品评论分析的整体结构

“应用层”：应用层的功能包含模型初始化、多条评论情感分析和多条评论关注点分析。应用层的逻辑包含两个部分，第一部分是进行数据可视化的功能，第二部分是对处理层进行调用、逻辑整合的功能，第一部分向第二部分发起功能请求，然后在第二部分将功能请求映射为某一组对于处理层的调用和整合。

“处理层”：处理层是情感分析应用的核心层，包含了文本预处理、文本对象空间模型建立、对象值计算、分类器训练、文本分析5大模块。处理的每个模块是提供给处理层的其它模块和应用层进行调用的，处理层的处理对象是数据层。

“数据层”包含两种文件，一种是用Mysql数据库存储的评论信息表、分词结果、词相似度计算结果、主题相关维度选择结果、文本模型结果、评论对象值计算结果、分类结果、关注点结果。一种是用文件形式存储的词集合文件、分类器模型文件、依存句法分析结果文件、HowNet文件、词向量模型文件、情感字典、词向量结果文件。

### 天猫商品评论分析应用的模块设计

按照图6-5所示的应用框架结构和6.2给出的各项功能的用例分析和行为分析，可以对处理层5个模块和应用层2个模块进行设计。

1. 处理层“文本预处理”模块由一个类Pretreatment构成，其中包含两个方法一个属性。属性是一个Comment实体类，这个类中只有4个属性，分别为：评论文本的编号、评论文本类型、评论文本字符串、所属主题。Pretreatment类的第一个方法cut()负责将属性Comment进行分词，结果写入分词结果表。Pretreatment类的第二个方法clear()负责将属性Comment的分词结果按词性划分写入词集合文件。
2. 处理层“文本对象空间模型建立”模块由4个类组成。



图6-6 文本对象空间模型建立类图



图6-7 文本对象空间模型建立时序图

其中Similarity类根据词集合计算词相似度，并将计算结果写入词相似度计算结果表；Object类根据词相似度计算结果表进行对象的抽取和归纳，将结果写入文本模型结果表；Subject类根据主题词在文本模型结果表中的名词对象中进行主题相关的维度选择；Create\_model类是对上述3个类的调用。

Similarity当中的属性SVM\_model是采用第4章实验过程中建立的分类法度量词相似度的SVM二分类器。

Create\_model类通过方法get\_Similarity()调用Similarity类的对应方法更新词相似度计算结果；通过方法get\_Object()调用Object类的对应方法更新文本模型结果；通过方法get\_Subject()调用Subject类的对应方法更新主题相关维度选择结果。

1. 处理层“对象值计算”模块由两个类组成，其中Chain类负责将评论短文本进行依存句法分析，得到以评论编号命名的分析结果文件；此外可以通过依存句法的分析结果文件得到词链，进而映射成为对象链；Value类负责调用Chain类的方法进行依存句法分析，然后根据分析结果文件得到词链，进而映射为对象链，然后查询情感字典，经过计算得到评论短文本的对象值，更新评论对象值计算结果。



图6-8 对象值计算类图

Value类通过方法get\_Chain()将属性调用Chain类的get\_Chain()方法，得到属性commentList当中所有的评论的对象链；通过方法get\_Value()根据评论的对象链得到评论的对象值，并写入到评论对象值计算结果。



图6-9 对象值计算时序图

1. 处理层“分类器训练”模块由一个类Classification构成，类包含的3个属性为BP网络分类器模型BP\_model、CHAID分类器模型CHAID\_model、训练数据集data，其中data是已经计算好对象值的已标注评论；2个方法为更新BP网络分类器模型的set\_BP()方法和更新CHAID分类器模型的set\_CHAID()方法。
2. 处理层“电子商务用户评论分析”模块由一个类Analysis构成，类包含的1个属性为data是已经计算好对象值的未标注评论；3个方法为利用BP分类器进行情感打分的get\_BP\_score()方法、利用CHAID分类器进行情感打分的get\_CHAID\_score()方法、统计data当中关注点和每个关注点的正负向评价比例的get\_Focus()方法。
3. 应用层“模型初始化”模块包含一个主类Init，类包含5个方法：init()进行初始化，对处理层的各个模块进行调用；pretreatment()调用“文本预处理模块”进行文本预处理；create\_model()调用“文本对象空间模型建立”模块进行文本对象空间的建立；value()调用“对象值计算”模块对已经存储的评论文本进行对象值计算，classification()调用“分类器训练”模块进行分类器的训练。整体的类图如图6-10所示。



图6-10 模型初始化类图

“模型初始化”模块的整体时序图如图6-11所示。



图6-11 模型初始化时序图

1. 应用层“多条评论情感分析”模块，由一个类Emotion构成，包含1个属性data，data是要进行情感评分的多条评论文本；包含一个方法emontion\_analysis()调用处理层的Analysis类中的利用BP分类器进行情感打分的get\_BP\_score()方法、利用CHAID分类器进行情感打分的get\_CHAID\_score()方法得到data的情感评分，并计算正负向评分比例。
2. 应用层“多条评论关注点分析”模块，由一个类Focus构成，包含1个属性data，data是要进行关注点分析的多条评论文本；包含一个方法focus\_analysis()调用处理层的Analysis类中统计关注点的get\_Focus()方法得到data的关注点和每个关注点的正负向评价比例。

以上论述了对天猫商城商品评论进行情感分析的应用中每个模块的设计和具体工作过程以及各个模块之间的关系。

## 天猫商品评论分析应用的模块实现

### 开发环境

本文采用Mysql和文件存储数据，采用j2SE开发处理层模块，通过JDBC和J2SE的文件访问功能与“数据层”进行交互，采用Java Servlet开发应用层的接受请求和返回请求，数据可视化采用Ajax+Echarts3。

表6-6 开发环境

|  |  |
| --- | --- |
| 功能模块 | 开发工具 |
| 数据可视化 | Ajax+Echarts3（WebStorm10.0.4） |
| 处理层开发 | Java（MyEclipse 10.7） |
| 数据层存储 | Mysql+文件 |

### 数据层实现

表6-7至表6-11给出一些关键数据表的实现。

分词结果表存储的是一条评论的分词结果。

表6-7 分词结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 备注 |
| comment\_id | varchar（255） | 评论唯一标识 | 外键 |
| comment\_type | varchar（255） | 评论的类型 | 外键 |
| comment\_cut | varchar（255） | 评论的分词结果 |  |
| subject\_id | varchar（255） | 所属主题 | 外键 |

词相似度计算结果表存储的是两个词的相似度结果，若词类型是名词，则字段is\_similar的值为1或者0；若词类型是情感词和程度词则给出字段similarity的值。

表6-8 词相似度计算结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 备注 |
| word\_1 | varchar（255） | 词语1 | 主键 |
| word\_2 | varchar（255） | 词语2 | 主键 |
| word\_type | int | 词类型 |  |
| similarity | varchar（255） | 相似度 |  |
| is\_similar | int | 词是否相似，1为相似，0为不相似 |  |
| from | int | 1为基于HowNet所得，0为基于词向量 |  |

文本模型结果表按对象标识存储每个对象包含的词语。

表6-9 文本模型结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 备注 |
| object\_id | varchar（255） | 对象标识 | 主键 |
| object\_type | int | 对象类型 |  |
| object\_include | varchar（255） | 包含的词语，词语之间以“/”隔开 |  |

主题相关维度选择结果表按主题标识存储每个主题相关的维度的对象标识。

表6-10 主题相关维度选择结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 备注 |
| subject\_id | varchar（255） | 主题标识 | 主键 |
| subject\_name | varchar（255） | 主题词 |  |
| include | varchar（255） | 包含的对象，对象之间以“/”隔开，每个对象包含对象ID和对象权值。如object1-0.2/object2-0.3/object2-0.5 |  |

评论对象值计算结果表存储的是每一条评论的对象值。

表6-11 评论对象值计算结果表

|  |  |  |  |
| --- | --- | --- | --- |
| 字段 | 类型 | 描述 | 备注 |
| comment\_id | varchar（255） | 评论标识 | 外键 |
| subject\_id | varchar（255） | 主题标识 | 外键 |
| value | varchar（255） | 根据主题标识得到对象维度，每个维度对象值之间以“/”隔开，如0.8/0.4/1.0 |  |

“依存句法分析结果文件”是按照comment\_id命名的\*.stp文件。词向量模型和分类器模型采用字符文件的形式存储参数。

### 处理层实现

根据6.3所述的各个模块的时序图和类图，可以给出各个模块的实现过程，通过进行每个模块实现过程的分析和一部分主要代码来给出。

1. 文本预处理模块的最重要的功能是完成评论文本的分词和清洗。为了解决分词的问题采用中科院的NLPIR分词工具。代码如下：

import com.sun.jna.\*;

public class Pretreatment {

Comment comment;

public interface CLibrary extends Library {}

public void cut(){

String system\_charset = "GBK";

int charset\_type = 1;

int init\_flag = CLibrary.Instance.NLPIR\_Init(argu

.getBytes(system\_charset), charset\_type, "0"

.getBytes(system\_charset));

String sInput =this.comment.comment\_text;

String nativeBytes = nativeBytes = CLibrary.Instance.

NLPIR\_ParagraphProcess(sInput, 3);

cut\_Results cut=new cut\_Results(Comment,nativeBytes);

cut.update();}

public clear(){

cut\_Results cut=new cut\_Results(Comment);

cut\_Results.clear();

}}

1. 文本对象空间模型建立模块当中最重要的功能是完成词相似度的计算、根据词相似度计算进行对象的抽取和归纳、在名词对象当中根据主题选择相关维度。词相似度的计算根据词集合的性质来进行区分计算，名词集合采用基于词向量的方法，情感词集合和程度词集合采用基于HowNet的方法；对象的抽取和归纳采用算法5.1；主题相关的维度的选择采用算法5.2。

首先是进行词相似度计算的类Similarity ，可以根据词集合的不同调用不同的方法进行比对。

public class Similarity {

wordFile wordFile;

SVM\_model svm;

String HowNetFile;

public void similarity\_n(){

//打开名词集合文件，并得到词向量

wordVec wv=new wordVec(wordFile.open(wordFile.n));

//根据词向量两两判定词的相似度，并存储

for(int i=0;i<wv.size-1;i++){

int svm\_result=svm.excute(wv.get(i),wv.get(i+1));

word\_similarity ws=new word\_similarity(svm\_result);

ws.update(wordFile.getword(i),wordFile.getword(i+1));

}}

public void similarity\_ad(){

for(int i=0;i<wordFile.ad.size;i++){

word\_similarity ws=new word\_similarity("HowNet");

New HowNet\_model(HowNetFile).excute(wordFile.getword(i),

wordFile.getword(i+1));

ws.update(wordFile.getword(i),wordFile.getword(i+1));

}}

}

然后是进行对象抽取和归纳的类WObject，相似的词放入同一个对象中。

public class WObject {

public void get\_Object(){

List<w\_object> ol=new ArrayList<w\_object>;

word\_similarity ws=new word\_similarity();

List<wordList> wl=word\_similarity.getList();

for(int i=0;i<wl.size();i++){

wordList one=wl.get(i)

if(one.word\_type==0){

if(one.is\_similar==1){

for(int j=0;j<ol.size();j++){

if(is\_similar(one,ol.get(j))){

w\_object.add(one);

}else{

ol.add(new w\_object(one));

}}

}}

if(one.word\_type==1){

if(one.similarity>=0.5){

for(int j=0;j<ol.size();j++){

if(is\_similar(one,ol.get(j))){

w\_object.add(one);

}else{

ol.add(new w\_object(one));

}}

}}}

new Object\_Model.update(ol);}}

然后进行主题相关的维度选择，根据主题词选择相关的对象维度。

public class Subject {

String TopicList;

public void ObjectSelect(){

List<Topic> tl=TopicFile.open(TopicList);

WObject o=new WObject();

Object\_Model OM=new Object\_Model();

for(int i=0;i<tl.size();i++){

Topic t=tl.get(i);

for(int j=0;j<t.size;j++){

for(int t=0;t<OM.size;t++){

if(o.is\_similar(t.getword(j),OM.get(t))){

t.set(i,t);

}}}}

new Topicrelated(tl);}}

最后是3个功能的整合调用类Create\_model。

public class Create\_model {

public init(){

this.get\_Similarity();

this.get\_Object();

this.get\_Subject();}

public get\_Similarity(){

new Similarity().similarity\_n();

new Similarity().similarity\_ad();}

public get\_Object(){

new WOject().get\_Object();}

public get\_Subject(){

new Subject.ObjectSelect();

}}

1. 对象值计算模块是根据预定义情感小字典对每一条评论的对象值进行计算。

public class Value {

List<Comment> commentList;

public void get\_Chain(){

for(int i=0;i<commentList.size();i++){

Chain chain=new Chain(commentList.get(i));

chain.get\_Chain();}}

public void get\_Value(){

for(int i=0;i<commentList.size();i++){

stp\_O\_List sol=new stp\_O\_List(commentList.get(i));

new commentValue().excute(sol);}

}}

public class Chain {

Comment Comment;

public void get\_Chain(){

LexicalizedParser.main(Comment.comment\_text);

Object\_Model OM=new Object\_Model();

String stpfile=Comment.commentid+".stp"

stpList sl=open(stpfile);

stp\_O\_List sol=sl.parser(OM);

sol.update();

}}

1. 分类器训练模块是利用已经计算好对象值的已标注评论进行BP分类器和CHAID分类器的训练。

public class Classification {

BP\_model BP\_model;

CHAID\_model CHAID\_model;

List<List<Double>> data;

public void set\_BP(){

this.BP\_model.train(data);}

public void set\_CHAID(){

this.CHAID\_model.train(data);

}}

1. 电子商务用户评论分析模块的主要功能是利用4中训练好的分类器对未标注评论进行分类的过程，过程比较简单，在此不做赘述。

综上所述，给出了处理层重要模块的功能实现分析和一部分的关键代码。

### 应用层实现

1. 模型初始化模块是调用处理层的各个模块，首先进行文本预处理，进而建立文本对象空间模型，然后计算每个评论的对象值，最后利用已标注的评论训练分类器。

public class Init {

public void init(){

this.pretreatment();

this.create\_model();

this.value();

this.classification();}

public void pretreatment(){

List<Comment> commentList=db.get\_CommentList();

for(int i=0;i<commentList.size();i++){

Pretreatment pr=new Pretreatment(commentList.get(i));

pr.cut();

pr.clear();}}

public void create\_model(){

Create\_model cm=new Create\_model();

cm.init();}

public void value(){

Value v=new value();

v.get\_Chain();

v.get\_Value();}

public void classification(){

List<List<Double>> data=db.get\_AlreadyFLAG()

Classification c=new Classification(data);

c.set\_BP();

c.set\_CHAID();

}}

2）应用层的“多条评论情感分析”模块和“多条评论关注点分析”模块是利用处理层更新的分类器直接对未进行类别标注的评论文本进行分类的过程，然后将分类的结果交给前端显示，过程较为简单，在此不做赘述。

## 天猫商品评论分析应用的模块测试

本章节对天猫商品评论分析应用的模块进行功能性测试。

### 测试环境

测试环境主要包括软件环境和硬件环境，软件环境是指被测试的软件运行时的操作系统、数据库、服务器等环境，硬件环境是指软件运行所需的硬件设备构成的环境。

表6-12 软件环境

|  |  |  |
| --- | --- | --- |
| 环境 | 名称 | 版本 |
| 操作系统 | Windows 7 | 64位专业版 |
| 浏览器 | chrome | 56.0.2924.87 (64-bit) |
| 数据库 | MySql | 5.1 |
| Web服务器 | Tomcat | 7.0 |

表6-13 硬件环境

|  |  |
| --- | --- |
| 环境 | 名称 |
| CPU | Intel(R) Core(TM) i5-4210M,2.6GHz |
| 内存 | 金士顿12GB |
| 网络是否连接 | 连接Internet网络 |

### 功能测试

测试是判断一个软件或者模块正确性、安全性、完整性的过程，其中黑盒测试也叫功能测试，只需要考虑测试各个模块和各个模块之间连接时的功能的正确性，即将预期输出和实际输出进行对比的过程。通过功能测试，能够确定各个模块的功能是否正常，从而可以确定整个情感分析模块的正确性。

功能测试的主要内容如下：测试各个模块自身功能是否符合需求；测试各个模块的执行流程是否符合设计当中的模块处理流程；测试集成后的运行，测试结果如下。

表6-14 词相似度计算测试结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | 用例编号 | 输入 | 预期结果 | 实际结果 | 结论 |
| 基于HowNet计算情感词相似度 | 01 | 3000个词对 | 得到3000个词对的相似度并更新词相似度结果文件 | 得到3000个词对的相似度并更新词相似度结果文件 | 通过 |
| 基于词向量计算名词相似度 | 02 | 6000个词对 | 得到6000个词对的相似度并更新词相似度结果文件 | 得到6000个词对的相似度并更新词相似度结果文件 | 通过 |

表6-15 文本对象空间模型建立模块测试结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | 用例编号 | 输入 | 预期结果 | 实际结果 | 结论 |
| 对象抽取 | 01 | 词相似度文件 | 得到抽取后的对象 | 得到抽取后的对象 | 通过 |
| 主题相关维度选择 | 02 | 5个主题 | 5个主题对应的子维度对象 | 5个主题对应的子维度对象 | 通过 |

表6-16 对象值计算模块测试结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | 用例编号 | 输入 | 预期结果 | 实际结果 | 结论 |
| 情感字典的访问 | 01 | 50个情感词 | 50个情感词值 | 50个情感词值 | 通过 |
| 对象情感值计算 | 02 | 50个对象链 | 50个文本对象情感值 | 50个文本对象情感值 | 通过 |

表 6-17 分类器训练模块测试结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | 用例编号 | 输入 | 预期结果 | 实际结果 | 结论 |
| BP分类器训练 | 01 | 5个主题，各自800条评论 | BP分类器训练完成，中间结果无误 | BP分类器训练完成，中间结果无误 | 通过 |
| CHAID分类器训练 | 02 | 5个主题，各自800条评论 | CHAID分类器训练完成，中间结果无误 | CHAID分类器训练完成，中间结果无误 | 通过 |

表6-18 文本分析模块测试结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 功能 | 用例编号 | 输入 | 预期结果 | 实际结果 | 结论 |
| 情感分类 | 01 | 5个主题，各自500条评论 | 给出每条评论的评分信息 | 给出每条评论的评分信息 | 通过 |
| 关注点计算 | 02 | 5个主题，各自500条评论 | 给出每个主题的关注点 | 给出每个主题的关注点 | 通过 |

通过上述的测试，本应用能够达到需求分析的要求，实现了对商品评论短文本进行文本对象空间建模、情感分析、多维度关注点分析等功能要求。

## 天猫商城用户评论分析结果

### 情感分析

本文的模型是基于根据情感词和程度词得到对象维度的值，所以可以用来做短文本的情感分析。将表5-3所列的天猫商城数据中11个大类做单条情感倾向分析，0.5分以下为负面评价，0.5分以上为正面评价。

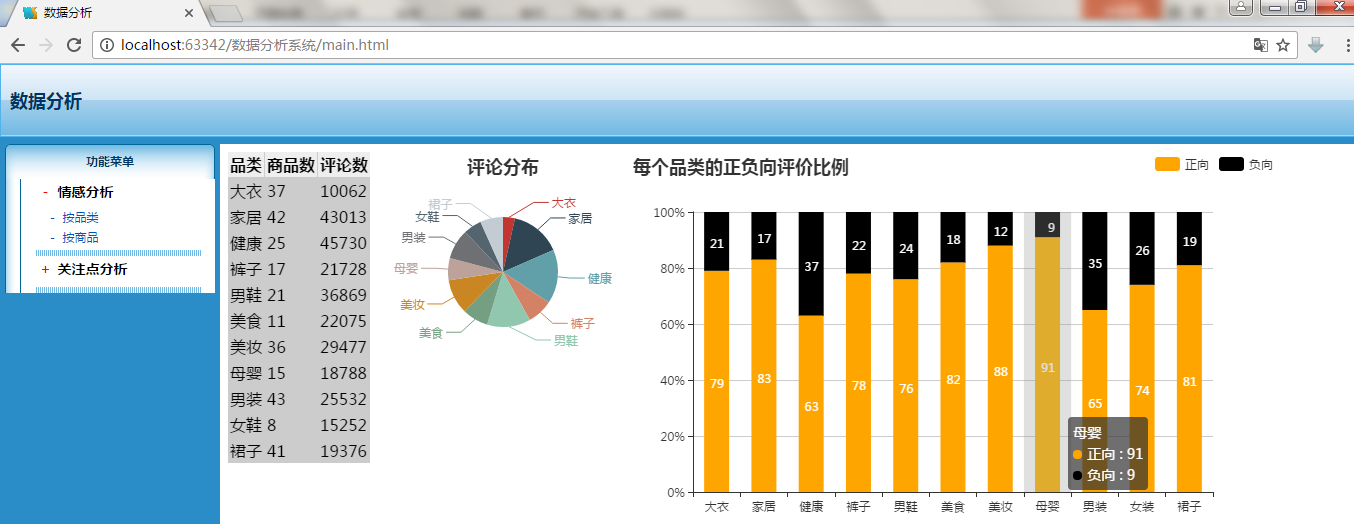


图6-12 按品类得到每个商品品类的正负面评价比例

页面的左侧给出分析的评论数据的数量和在品类当中的分布，页面右侧给出每个品类的正面评价和负面评价的比例。可见，在天猫商城的评论文本当中包含着大量的负面评论，而这些负面评论都被天猫商城评论系统所固有的问题所掩盖而不能真实的呈现给普通消费者。

其中“健康”品类的负面评价比例相对较高，这其实是符合常识的，很多健康用品并不如其宣传的那么神效，当普通消费者买到商品之后发现了其虚假和浮夸的地方，自然会通过对商品的评论去体现意图，卖家在这种时候往往利用天猫商城评论系统当中买家可以二次改分的行为性漏洞，通过各种方式去请求或者骚扰消费者改掉不良的分数，从而掩盖了商品的问题。

将表5-3所列的天猫商城数据中再随机抽取11个商品的评论采用分类法做按商品评分，同已有评分做对比。这种对比方式可以得到每一个商品的评论的真实评分和天猫商城评论系统所带来的评分之间的差异性，这样的差异性可以体现出天猫商城评论系统的问题，同时可以使得消费者在搜索某个商品的时候，可以通过这个功能准确的得到某个商品的真实评分，从而不被天猫商城的评论系统的问题所误导。页面上通过选择商品的品类和商品的名称，可以给出本文的模型评分和已有评分的对比。

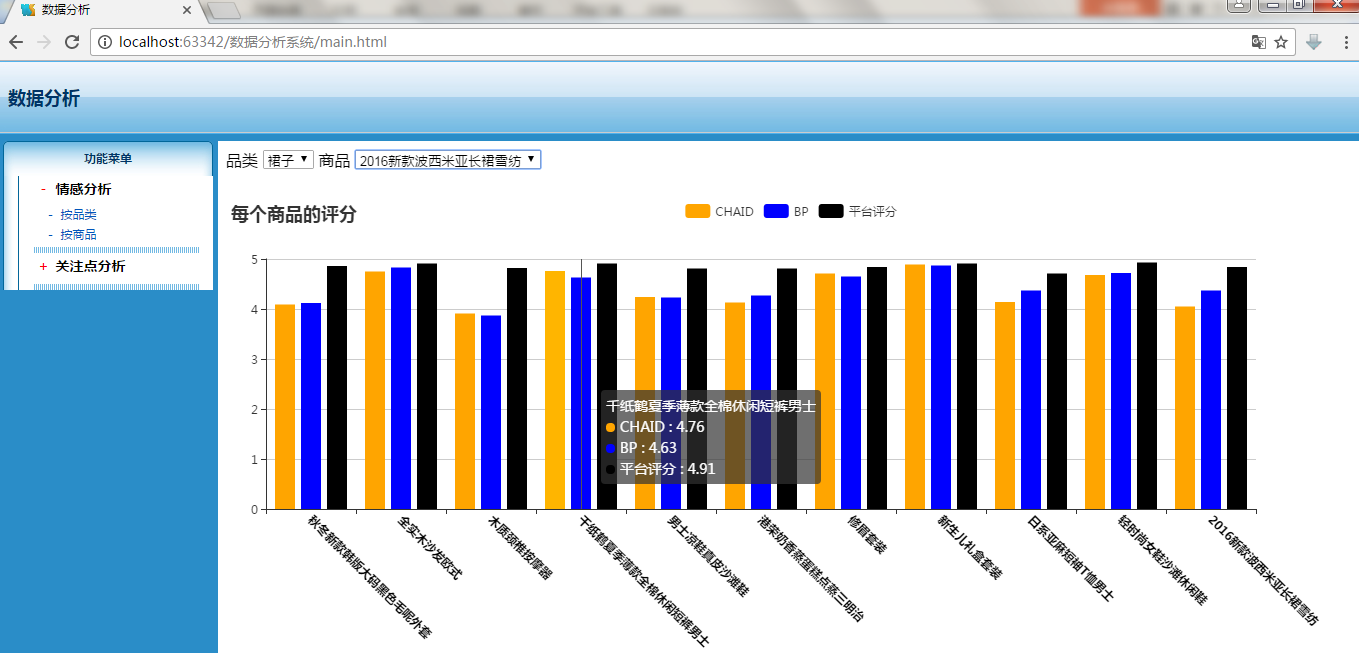


图6-13 选择商品，显示其评分

从图6-13可以看出，利用模型得到的CHAID评分和BP评分都比已有的商品评分要低，这并不是说明模型错误，而是说明天猫用户的打分与其评论的真实意思有一定的差距，有的用户的打分并不能准确的反映用户的真实评价，用户的打分受到多重影响，比如卖家回电要求改分等情况。

### 关注点分析

文本建立的文本对象空间模型，能够在对象维度上反映主题相关的属性，维度的值可以反映评论文本反映的态度，这样就可以利用本文的模型进行商品用户关注点分析。在将表5-3所列的天猫商城数据中11个大类利用统计的方法进行关注点提取，并给出每个关注点的好评和差评的比例。

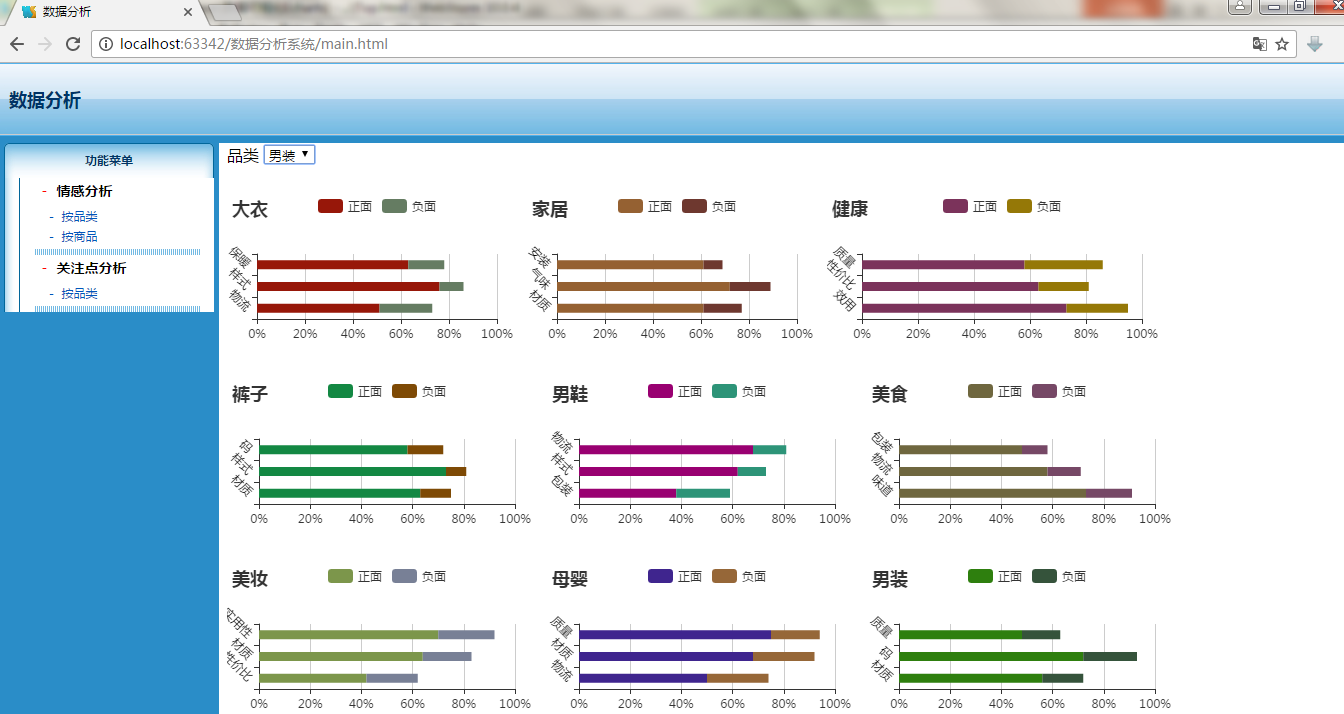


图6-14 品类关注点及正负面评价比例

页面通过选择品类来呈现每个品类的3个关注点，并且给出每个关注点的正面评价与负面评价的比例，通过利用本文提出的模型进行关注点分析，同时加入每个关注点的情感分析，就可以知道用户关注什么、商品本身的优缺点是什么。

## 本章小结

本章首先分析了天猫商城评论系统的特点和其存在的问题，分析了这些问题对普通消费者造成的影响，提出了解决问题的目标，然后简要介绍了基于本文提出的文本对象空间模型对天猫商城评论数据进行分析的设计与实现过程，给出了框架结构和一些核心的功能运行过程，最后给出了天猫商城用户评论的分析结果。

# 结论与展望

## 结论

本文聚焦于短文本情感分析当中降维方法、维度选择和情感倾向计算的研究与应用，首先建立一个全局的文本对象模型，然后在此模型当中根据短文本本身的特性进行属性选择和表示。本文提出了文本对象空间模型，认为其中的关键技术在于利用词相似度的计算进行对象的抽取和归纳。本文提出了基于HowNet的改进的词相似度计算方法和基于改进的CBOW产生的词向量进行词相似性度量的方法，并将两种计算方法应用于不同的词性的对象抽取和归纳，在最终形成的文本对象空间模型当中进行文本分析。现将本文的主要工作总结如下：

1）本文深入研究了短文本的特点和经典的向量空间模型的优点，针对短文本，引入词汇链模型和布尔模型的特性，在文本词空间上利用降维形成文本对象空间模型。

2）本文针对词维度的降维提出了基于词相似度进行降维，将一堆相似的词看做同一个对象。本文改进后的基于HowNet的词相似度计算方法的准确率由0.713提升到了0.842，其中情感词和程度词的相似度准确率提高到了0.981和0.965。为了能够更好的处理未收录词，对神经概率模型当中的CBOW模型进行了改进。实验表明，改进模型将词余弦相似度的准确率提升到0.764，其中名词的余弦相似度准确率提高到了0.872。本文进而针对词的向量化提出的分类度量法能够比距离度量法更加准确和全面表达词相似性，同时在计算难度上优于相异性度量法，并且能够在无需指定相似度阈值的情况下将词相似度的准确率提高到0.887。

3）本文选择基于HowNet的词相似度计算方法针对情感类词汇进行对象抽取和归纳，选择基于改进的CBOW模型的词向量对名词类词汇进行处理。根据短文本的句子成分依存关系建立对象链关系，利用对象链计算每个对象维度的情感值。建立了主题相关的维度选择，在整个文本对象空间当中进行与分析主题相关的维度选择，能够更加精确的表示短文本。

4）通过抓取的亚马逊商城的的评论文本和评分信息进行实验，实验表明，本文提出的文本对象空间模型对电子商务用户评论短文本进行情感评分时，在主题范围比较明确的品类下准确率可以达到0.934，平均准确率可以达到0.847，能够很好的表达短文本所描述的事物和情感倾向。

5）本文最后基于本文提出的文本对象空间模型开发了一个面向天猫商城商品用户评论分析的小应用，展示了天猫商城商品用户评论情感分析和关注点分析的结果。

## 展望

本文围绕评论短文本的情感分析进行了深入研究，包括词相似度计算方法和文本的模型化表示，取得了阶段性的成果。但是本文提出的方法仍需要进一步的拓展。

1）本文提出的基于HowNet的词相似度计算当中，相同词性概念和不同词性概念的权值应该是动态变化，在不同的词当中应当具有不同的权值分布。相应的在概念相似度计算当中，本文是根据义项表达式中的义原的位置权值，下一步可以考虑是否可以利用义项表达式在义原层次树中的子树形态进行概念相似度的比较，需要进一步进行研究。

2）本文提出的改进的CBOW模型在投影层的处理时输入层的加权求和，每个输入单元词的权值是根据与输出词之间的物理距离来计算的，下一步考虑是否可以利用句子成份依存关系来进行权值计算，有关系的词的权值应该增加，需要进一步的实验验证。

3）本文在进行对象值计算的时候采用预定义部分字典的形式来给出情感倾向的度量，这种方法属于人工干预，下一步可以研究能否自动化的进行情感倾向值的赋予，在此基础上可以改进本文所给出的计算方法。

# 致 谢

在此谨向我在研究生学习期间给予我帮助的人们表达诚挚的谢意。这三年来我的学习和生活离不开你们的支持。

首先，感谢我的导师曾明教授。在这三年的中无论是学习还是生活方面对我关怀备至，在研究方面给予我指导和建议，每一次和曾明老师谈话都感触良多。曾老师对于工作一丝不苟，对于生活热情如火，这些都深深的影响到了我。在此，向曾老师致以衷心的感谢和诚挚的敬意。

其次，感谢我们的大师兄曾轲，在多次的碰面和小组会议当中，他指导我如何去研究，如何去发现新的思路，对我的学习有极大的帮助，对我论文的完成指明了具体方向和方法。再次谢谢曾轲师兄，并衷心的祝福他在工作和生活当中一切顺利。

再次，感谢所有软件学院的老师，你们的授课让我获益匪浅，感谢我的同学们，同你们的一起的生活和学习、和你们的一起的讨论，使我取得了长足的进步。

最后，感谢我的家人，父母对我生活的无微不至的关心，妻子对我的爱与理解，有你们，我才能顺利的完成学业，在此，对他们致以深深的谢意，愿与他们分享这份快乐与喜悦。

# 

# 参考文献

1. CNNIC.第39次中国互联网络发展状况统计报告[EB/OL]. : 127(2017).

http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201701/t20170122\_66437.htm．

1. 陈现麟，王樊．利用上下位关系的中文短文本分类[J].计算机应用,2010,30(3):603-606.
2. 孙建旺，吕学强，张雷瀚.基于语义与最大匹配度的短文本分类研究[J].计算机工程与设计， 2013,34(10)：3613-3618.
3. 詹志建，梁丽娜，杨小平．基于百度百科的词语相似度计算[J]．计算机科学，2013,40(6)： 199-202.
4. Wang F,Wang Z,Li Z,et al.Concept-based Short text classification and ranking[C]//Proceeding of the 23rd ACM International Conference on Information and Knowledge Management.AC

M,2014:1069-1078．

1. Kim K,Chung B,Choi Y,et al.Language independent semantic kernels for short-text classification[J].Expert Systems with Applications,2014,41(2):735-743.
2. Meng W,Lanfen L,Jing W,et al.Improving short text classification using public search engines[M]//Integrated Uncertainty in Knowledge Modelling and Decision Making.Springer Berlin Heidelberg,2013:157-166.
3. Dai Z,Sun A,Liu X Y.Crest:Cluster-based Representation Enrichment for Short Text Classification[M]//Advances in Knowledge Discovery and Data Mining.Springer Berlin Heidelberg,2013:256-267．
4. Chen M,Jin X,Shen D.Short text classification improved by learning multigranularity topics[C]//IJCAI.2011:1776-1782.
5. 刘帅，凌剑勇，吴元昊．基于条件随机场的产品特征提取方法[J].电脑知识与技术， 2016,12(3):187-190.
6. Jaap kamps.Words with attitude[M].In Proceedings of the 1st International Conference on Global WordNet.2002:332-341.
7. Wiebe.Word sence and Subjectivity[M].In Proceedings of the 21st ICCL and the 44th annual meeting of the ACL2006:1065-1072.
8. 朱嫣岚．基于HowNet的词汇语义倾向研究[J]．中文信息学报，2006,20(1):14-20.
9. 孙莹．基于Web文本挖掘的企业口碑情感分类模型研究[D].武汉：华中师范大学，2013．
10. Agiree E,Rigau G.A proposal for word sense disambiguation using conceptualdistance[C]//In-

ternational Conference/Recent Advances in Natural Language RecessingRANLP.95.Tzigov-

Chark,Bulgaria,1995:91-98.

1. Brown P.Word sense disambiguation using tactical methods[C]//Proceedings of the 29th meeting of the Association for Computational Linguistics（ACL291）.1991:210-207.
2. Li Su-jian,et al.Semantic computation in Chinese question-answering system[J].Journal of Computer Science and Technology,2002,17(6)：933-939.
3. 李茹，王智强，李双红．基于框架语义分析的汉语句子相似度计算[J].计算机研究与发展， 2013,50(8):1728-1736.
4. 詹志建，杨小平．基于语言网络和语义信息的文本相似度计算[J].计算机工程与应用， 2014,20,(5):33-38．
5. SANGEETHA S,THAKUR R S,AROCK M. Event detection using lexical chain[C]//LOFTSS

ONH. Advances in natural language processing. Berlin: Springer Berlin Heidelberg,2010:31

4-319.

1. 数据挖掘导论[DB/OL]. http://book.csdn.net/bookfiles/327Introduction of data mining[DB/OL]. http://book.csdn.net/ bookfiles/327．
2. 董振东，董强，知网简介[EB/OL].http://www.keenage.com/html/c\_index.html.
3. Bengio Y,Schwenk H,Senecal J S,et al.Neural probabilistic language models[M]//Innovations in Machine Learning.Springer Berlin Heidelberg,2006:137-186.
4. Mikolov T,Karafiat M,Berget L,et al .Recurrent neural network based language model[C]//INTER-SPEECH.2010:1045-1048.
5. Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word Representations in Vector Space[EB/OL].http://arxiv.org/abs/1301.3781,2013-01-16.
6. 刘群，李素建．基于《知网》的词汇语义相似度计算[J]．中文计算语言学,2002,7(2)：59-76．
7. 李峰，李芳.中文词语语义相似度计算-基于《知网》2000[J].中文信息学报， 2007,21(3):99-105．
8. 张沪寅，刘道波，温春艳．基于《 知网》的词语语义相似度改进算法研究[J]．计 算 机 工 程，2015,41(2)：151-156.
9. 朱征宇，孙俊华．改进的基于《知网》的词汇语义相似度计算[J]．计算机应用， 2013,33(8):2276-2279．
10. Wang Xiao Lin,Yang Lin.New word similarity algorithm research based on HowNet[J].Information Science,2015,2:67-71.
11. Lin Dekang.Automatic Retrival and Clustering of Similar words[C]//Proceedings of COLING-ACL’98.Montreal,Canada:[s.n.],1998:768-774.
12. 盛雅琦,张晗,吕晨,等．基于混合主题模型的文本蕴涵识别[J]．计算机工程， 2015,41( 5):180-184.
13. 王桐,王磊,吴吉义,等．WordNet中的综合概念语义相似度计算方法[J].北京邮电大学学报， 2013,36(2):98-101.
14. 罗杰，王庆林，李原．基于Word2vec与语义相似度的邻域词聚类[C]//第三十三届中国控 制会议论文集．北京：中国学术期刊电子杂志出版社，2014:517-522.
15. 郑文超，徐鹏．利用word2vec对中文词进行聚类的研究[J].软件，2013(12):160-162.
16. 鉴萍．依存句法分析方法研究与系统实现[D]．北京：中国科学院自动化研究所，2010.
17. Lucia W,Ferrari E.EgoCentric:Ego Networks for Knowledge-based Short Text Classification[C]//Proceeding of the 23rd ACM International Conference on Information and Knowledge Management.ACM,2014:1079-1088．
18. Remus S,Biemann C.Three knowledge-free methods for automatic lexical chain extraction[C]//Proceedings of NAACL-HLT 2013.Stroudsburg:Association for Computational Linguistics,2013:989-999.
19. 易寒飞．汉语文本语义情感分析研究[D]．北京：北京航空航天大学，2010．
20. 汪祥，贾焰，周斌，等．基于中文维基百科链接结构与分类体系的语义相关度计算[J]．小 型微型计算机系统，2011,32(11):2237-2242．
21. 孙琛琛，申德荣，单菁，等．一种基于维基百科结构信息的语义关联度计算算法[J]．计 算机学报，2012,35(11):2361-2370．
22. 黄承慧，印鉴，候昉，等．一众结合词项语义信息和TF-IDF方法的文本相似度量方法[J]．计 算机学报，2011(5):856-864．
23. 覃世安，李法运．文本分类中TF-IDF方法的改进研究[J]．现代图书情报技术， 2013,238(10):27-30．
24. 刘志明，刘鲁．基于机器学习的中文微博情感分类实证研究[J]．计算机工程与应用， 2012,48(1):1-4．

学位论文独创性声明（1）

本人声明：所呈交的学位论文系在导师指导下本人独立完成的研究成果。文中依法引用他人的成果，均已做出明确标注或得到许可。论文内容未包含法律意义上已属于他人的任何形式的研究成果，也不包含本人已用于其他学位申请的论文或成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．交回学校授予的学位证书；

2．学校可在相关媒体上对作者本人的行为进行通报；

3．本人按照学校规定的方式，对因不当取得学位给学校造成的名誉损害，进行公开道歉。

4．本人负责因论文成果不实产生的法律纠纷。

论文作者（签名）： 日期： 年 月 日

学位论文独创性声明（2）

本人声明：研究生 所提交的本篇学位论文已经本人审阅，确系在本人指导下由该生独立完成的研究成果。

本人如违反上述声明，愿意承担以下责任和后果：

1．学校可在相关媒体上对本人的失察行为进行通报；

2．本人按照学校规定的方式，对因失察给学校造成的名誉损害，进行公开道歉。

3．本人接受学校按照有关规定做出的任何处理。

指导教师（签名）：日期： 年 月 日

学位论文知识产权权属声明

我们声明，我们提交的学位论文及相关的职务作品，知识产权归属学校。学校享有以任何方式发表、复制、公开阅览、借阅以及申请专利等权利。学位论文作者离校后，或学位论文导师因故离校后，发表或使用学位论文或与该论文直接相关的学术论文或成果时，署名单位仍然为西安交通大学。

论文作者（签名）： 日期： 年 月 日

指导教师（签名）： 日期： 年 月 日

(本声明的版权归西安交通大学所有，未经许可，任何单位及任何个人不得擅自使用)